Editores(s): Márcio Dorn e Luis C. Lamb

REVISTA DE INFORMÁTICA TEÓRICA E APLICADA

VOL. 23, N. 1, Ano: 2016

29 de maio de 2016

Insituto de Informática
Universidade Federal do Rio Grande do Sul

REVISTA DE INFORMÁTICA TEÓRICA E APLICADA

VOL. 23 N. 1, Ano: 2016

CIP - Catalogação na publicação

Revista de Informática Teórica e Aplicada - Vol. 23, n. 1 -.- Porto Alegre: Instituto de Informática UFRGS, 2016.

Duas edições por ano.
Disponível online em: http://seer.ufrgs.br/rita
Descrição baseada em: Vol. 23, n. 1 (Mai. 2016).
ISSN 2175-2745

1. Informática.

Arte do Miolo: Alfeu Tavares

Revista de Informática Teórica e Aplicada - RITA

A Revista de Informática Teórica e Aplicada - RITA, é editada sob a responsabilidade do Instituto de Informática da Universidade Federal do Rio Grande do Sul, e visa publicar trabalhos que mostrem o estado da arte e tendências da área de Informática e suas aplicações, servindo também como um fórum para discussões de projetos em desenvolvimento nas universidades e centros de pesquisa, bem como de seus resultados e perspectivas de aplicação. Com periodicidade normal semestral, a RITA também serve como veículo de divulgação de trabalhos selecionados em eventos através de edições especiais. Sua abrangência é, basicamente, a comunidade Íbero-latino americana, com matérias aceitas em português, espanhol e inglês.

Política de Acesso Livre: Esta revista oferece acesso livre imediato ao seu conteúdo, seguindo o princípio de que disponibilizar gratuitamente o conhecimento científico ao público proporciona maior democratização mundial do conhecimento.

Principais áreas: A revista RITA publica artigos nos seguintes campos investigativos:

- Bioinformática
- Computação de Alto Desempenho
- Computação Gráfica
- Engenharia de Software
- Fundamentos da Computação e Métodos Formais
- Informática na Medicina
- Inteligência Artificial
- Microeletrônica, Concepção de Circuitos Integrados
- Nanotecnologia Computacional e Nanocomputação
- Otimização
- Redes de Computadores
- Robótica
- Sistemas de Informação
- Sistemas Embarcados

Assinaturas: A revista RITA passou, a partir de 2010, a ser editada apenas em sua versão eletrônica, ISSN 2175-2745, com três edições por ano.

SUMÁRIO:

Linguagens de consulta para bases de dados em grafos: um mapeamento sistemático

Query languages for graph databases: a systematic mapping

Simone de Oliveira Santos [1]
Martin A. Musicante [2]
Mirian Halfeld Ferrari Alves [3]

Data de submissão: 09/03/2015, *Data de aceite: 04/05/2016*

Resumo: A popularização das redes sociais, associada à necessidade de analisar e sumarizar grandes volumes de dados oriundos das mesmas tem favorecido o uso de bases de dados em grafos. As linguagens de consulta a este tipo de base de dados devem, ao mesmo tempo, ter expressividade suficiente para a realização de consultas complexas e possibilitar o processamento eficiente de grandes volumes de dados. Este artigo apresenta um mapeamento sistemático sobre as linguagens de consulta para bases de dados em grafos, com foco nas suas características principais como paradigma ou capacidade de agregação de dados. O foco deste mapeamento é investigar e quantificar as publicações referentes às linguagens de consulta, caracterizando-as, identificando possíveis áreas de pesquisa, tendências e desafios.

***P*alavras-chave:** bases de dados em grafos, linguagem de consulta, mapeamento sistemático

Abstract: The widespread use of digital social networks, together with the need for the analysis and summarisation of their contents is at the base of the widespread use of graph databases. Query languages for this kind of databases need to be expressive enough to express complex queries to be efficiently performed over enormous amounts of data. In this paper, we perform a systematic mapping about query languages for graph databases, covering aspects such as their paradigm or aggregation capabilities. We focus on the quantitative analysis of publications on the subject, to better identify trends and future challenges.

***K*eywords:** graph database, query language, systematic mapping

[1]PPgSC Programa de Pós-graduação em Sistemas e Computação, UFRN Universidade Federal do Rio Grande do Norte - Natal, Rio Grande do Norte, Brasil. {simone82@ppgsc.ufrn.br,mam@dimap.ufrn.br}
[2]PPgSC - Programa de Pós-graduação em Sistemas e Computação, DIMAp Departamento de Informática e Matemática aplicada, UFRN Universidade Federal do Rio Grande do Norte - Natal, Rio Grande do Norte, Brasil
[3]Université d'Orléans, INSA CVL, LIFO EA - Orléans, France. {mirian@univ-orleans.fr}

1 Introdução

Grafos são populares para representar dados. Eles podem ser usados para modelar relacionamentos entre objetos e recursos. Grafos podem ser definidos como uma coleção de vértices e arestas, onde os vértices representam entidades, e as arestas representam a forma como as entidades se relacionam [78]. As bases de dados em grafos fazem parte de um conjunto de tecnologias criadas ou redescobertas para suprir algumas limitações do modelo de dados relacional [1].

O modelo de dados em grafo pode ser definido como aquele em que tanto o esquema quanto as instâncias são modelados como grafos [3]. Nos últimos anos, este modelo de dados tem recebido muita atenção da indústria e da academia. Com o avanço da Internet, o compartilhamento de dados se tornou um tópico importante e uma grande quantidade de grafos está disponível em bases de dados para uso experimental [12]. Os grandes grafos ganharam destaque e cada vez mais aplicações são feitas para lidar com esse tipo de volume.

É possível notar o crescente interesse pelas bases de dados em grafos pela literatura, inclusive conferências específicas dedicadas ao assunto[4]. Estudos relacionados às bases de dados em grafos podem ser direcionados para a forma de armazenamento dos dados (estrutura dos dados), para a forma de manipulação desses dados através de uma linguagem de consulta e também nas restrições de integridade apropriadas para o contexto.

Este artigo apresenta um mapeamento sistemático [74] sobre as linguagens de consulta para bases de dados em grafos. Para atingir esse objetivo, também será investigado o retorno ao uso do modelo de dados em grafos, apresentando os números referentes a este tópico e as razões para este comportamento. Esta investigação envolve identificar o tipo de contribuição que os artigos se propõem, como também que tipo de problema eles tentam resolver, e que técnicas eles usam para atingir os seus objetivos.

Um mapeamento sistemático é um método para analisar um tópico de interesse seguindo passos bem definidos. Ao final da análise, deve-se responder um conjunto de questões de pesquisa que são lançadas no início do processo. Toda a investigação dos números do mapeamento devem servir de base para responder as questões lançadas. O objetivo deste trabalho é investigar e quantificar as publicações referentes às linguagens de consulta, caracterizando-as, identificando possíveis áreas de pesquisa, tendências e desafios.

Este artigo é organizado como segue. A Seção 2 apresenta conceitos referentes ao método usado para a confecção deste mapeamento sistemático, e considerações referentes ao modelo de dados em grafos e trabalhos relacionados. A Seção 3 apresenta todo o processo de mapeamento, etapa a etapa. A Seção 4 apresenta as análises quantitativas extraídas do

[4]Por exemplo, as conferências GraphConnect (http://graphconnect.com/) e o International Workshop on Querying Graph Structured Data (http://www.isgroup.unimo.it/graphq2016/).

mapeamento e a Seção 5 contém as conclusões finais.

2 Revisão Teórica

Esta seção é dedicada a apresentar o processo de mapeamento sistemático que será aplicado neste trabalho e apresentar conceitos referentes ao modelo de dados em grafos e linguagens de consulta.

2.1 Mapeamento sistemático

Um mapeamento sistemático é um tipo de estudo que fornece uma visão geral sobre uma área específica de pesquisa [74]. De forma geral, o processo é feito coletando artigos referentes a um assunto, publicados em uma época determinada. A partir dos dados coletados é realizado um mapeamento entre os diferentes aspectos dos artigos coletados, de forma a poder fazer uma análise quantitativa do conjunto, verificando relações e tendências. O mapeamento é feito geralmente a partir da leitura dos títulos e resumos dos artigos. A análise quantitativa é ilustrada com gráficos e tabelas. O método utilizado neste trabalho foi feito de acordo com [74], que contém cinco grandes etapas, como mostrado na Figura 1, e cada etapa será descrita a seguir.

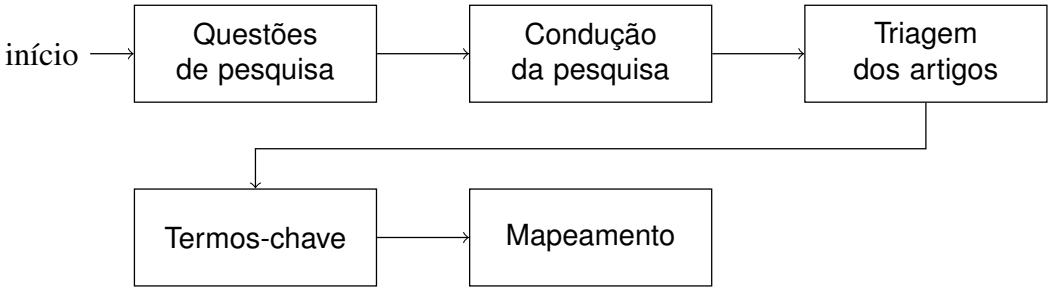


Figura 1. Etapas do mapeamento sistemático aplicadas neste trabalho.

2.1.1 Questões de pesquisa A primeira parte do mapeamento consiste em definir o objetivo principal do mesmo. As questões de pesquisa definem o assunto e o objetivo do mapeamento. As questões indagam sobre o assunto escolhido e indicam quais são os interesses específicos. As questões de pesquisa são importantes também para definir quais artigos devem fazer parte do mapeamento.

2.1.2 Condução da pesquisa Este passo refere-se ao processo de coletar artigos nas bases de dados que reúnem as publicações de conferências e periódicos científicos. A condução da

pesquisa consiste em três etapas: definição dos termos de consulta; escolha das bases de dados; e a execução da consulta nas bases de dados.

A definição dos termos de consulta refere-se à criação de palavras ou termos (*string* de busca) usados para consultar os artigos nas bases de dados. A escolha das bases de dados deve levar em conta as bibliotecas mais populares e abrangentes. Atualmente as consultas geralmente são feitas em bibliotecas digitais *on-line* que reúnem publicações das principais conferências e periódicos científicos. A execução da consulta nas bases de dados consiste em utilizar a *string* de busca nas bibliotecas digitais e recuperar todos os artigos que casaram com a consulta feita.

2.1.3 Triagem dos artigos A fase de triagem é feita após a recuperação dos artigos das bases de dados de bibliotecas digitais previamente selecionadas. Os artigos recuperados são organizados em listas e a triagem serve para escolher quais artigos devem ser selecionados para fazer parte do mapeamento ou não. A triagem é feita com a leitura dos títulos e resumos dos artigos somente, com base em critérios de inclusão e exclusão de artigos.

Os critérios de inclusão e exclusão são definidos baseando-se nos objetivos do mapeamento, principalmente usando as questões de pesquisa. Os critérios definem quais artigos são relevantes ou não para o mapeamento. É importante incluir nos critérios de exclusão a verificação de artigos repetidos, haja vista que um mesmo artigo pode estar incluído em mais de uma base de dados.

2.1.4 Termos-chave A busca por termos-chave é usada para criar uma classificação dos artigos. Os termos-chave são identificados durante a leitura dos títulos e resumos dos artigos. Estes termos são aqueles que identificam a contribuição do trabalho, pontos importantes como o objetivo, os métodos e tecnologias usados. A classificação neste contexto, é o ato de combinar os termos identificados e definir um conjunto de categorias. A divisão em categorias é importante e ajuda no momento de fazer uma análise e tirar conclusões. Os artigos selecionados para o mapeamento devem ser distribuídos dentro das categorias definidas.

2.1.5 Mapeamento Nesta etapa os artigos devem estar distribuídos entre as categorias definidas. A classificação não precisa ser estática, ela pode evoluir por exemplo com a criação ou mesclagem de categorias. Esta é a fase onde o mapeamento toma forma e as análises são feitas. A análise dos resultados deve ter como foco o número de publicações em cada categoria. Isso permite visualizar a progressão das publicações, como elas se distribuem aos longo dos anos, ou se algum assunto está em ênfase ou caindo em desuso. Neste ponto, também é oportuno identificar se há uma relação entre as categorias, e ilustrá-las se for o caso. Ao final da análise é importante que as questões de pesquisa sejam respondidas com base nos resultados obtidos com o mapeamento.

2.2 Modelo de dados em grafos e linguagens de consulta

Modelos de dados são usados para representar a forma ou esquema dos dados de uma organização. Um modelo de dados consiste em (i) um conjunto de tipos estruturados de dados, (ii) um conjunto de operadores ou regras de inferências, e (iii) um conjunto de regras de integridade [1]. Este trabalho tem maior foco no conjunto de operadores, ou seja, nas linguagens de consulta que são usadas para manipular os dados. O uso das bases de dados em grafos tem captado o interesse da indústria e academia na última década [101]. É possível observar um aumento no número de artigos publicados neste contexto a partir da década de 2000. Este comportamento pode ser observado na Seção 4.

O surgimento de novas aplicações relacionadas com a Internet, como as redes sociais digitais e a web semântica, demandaram novas tecnologias. Dentre estas novas tecnologias, o *Resource Description Framework* (RDF) [22] surgiu como um formato para representar metadados na *Web*, mas atualmente ele é bastante usado como uma extensão do modelo de dados em grafos [2]. Um documento do tipo RDF é uma coleção de triplas no padrão *sujeito-predicado-objeto* representando, dessa forma, os arcos de um grafo (o predicado), ligando um par de nodos (sujeito e objeto). Este tipo de grafo, que é considerado por alguns como um modelo de dados, é uma das tecnologias usadas na web semântica que provê um ambiente onde aplicações podem manipular (consultar/alterar) dados e fazer inferências [86]. A linguagem de consulta SPARQL é a linguagem recomendada pela W3C[8] para grafos RDF, mas de forma geral, não há uma linguagem de consulta padrão para bases de dados em grafos [12].

Outro tipo de tecnologia voltada para o modelo de dados em grafos são os sistemas nativos. Os sistemas nativos consideram a estrutura de grafos tanto no armazenamento físico dos dados quanto no processamento de consultas [72]. Estes sistemas são feitos especificamente para gerenciar grafos e usam listas de adjacência para processá-los. Um exemplo para esse tipo de sistema é o Neo4j [5], que utiliza a linguagem de consulta Cypher, que também dá suporte ao SPARQL.

A agregação é uma função bastante usada no modelo de dados relacional. Uma função de agregação usa o resultado de uma consulta como valor intermediário para fazer uma outra operação. As funções de agregação mais comuns são soma, média, maior valor e menor valor. O uso desse tipo de função nas linguagens de consulta é observado neste mapeamento.

2.3 Trabalhos relacionados

Existem vários trabalhos que fazem comparativos, ilustram o estado-da-arte, ou fazem revisões da literatura dedicados às bases de dados em grafos. Esses trabalhos normal-

[5] http://neo4j.com/

mente exploram os diversos modelos de dados e diferentes tipos de aplicações. Alguns deles dedicam-se exclusivamente às linguagens de consulta para grafos. Peter Wood mostra em [101] uma visão geral das linguagens de consulta para grafos desenvolvidos nos últimos 25 anos. O autor explora diferentes sintaxes e aspectos das funcionalidades.

O *survey* [5] tem como objetivo apresentar linguagens de consulta propostos para os principais formalismos da web semântica, tais como RDF e outros. O artigo discorre sobre várias linguagens e tenta destacar importantes aspectos sobre elas. Os autores Angles e Gutierrez escreveram o *survey* [3] que é uma referência para o modelo de dados em grafos. Eles dedicam uma seção para descrever as linguagens de consulta para grafos que servem como referência na área.

Este presente mapeamento pode ser visto como um complemento aos *surveys* já existentes e com o diferencial de ser apresentado em um formato (mapeamento sistemático) que não foi encontrado abordando o assunto em questão. Um mapeamento tem também como objetivo quantificar as publicações encontradas em um determinado período, dando a oportunidade de fazer conjecturas baseando-se nestes números. A maior diferença entre os *surveys* e o mapeamento sistemático está no fato de que o mapeamento dá uma ênfase maior nos números de publicações, divisão em grupos, análise quantitativa.

3 Processo de mapeamento sistemático

A execução deste mapeamento foi feita seguindo o método proposto em [74], o qual foi apresentado na Seção 2.1. Nesta seção, será descrito em detalhes como foi desempenhada cada etapa do processo de mapeamento. A última etapa, referente à extração de dados e mapeamento será apresentada na Seção 4.

3.1 Questões de pesquisa

O objetivo deste mapeamento sistemático é investigar quais linguagens de consulta para bases de dados em grafos são mais usadas e quais suas principais particularidades. Características das linguagens como a agregação estão entre as contribuições buscadas e ainda, quais paradigmas são mais usados para a construção dessas linguagens.

Baseado neste objetivo foram definidas três questões de pesquisa para a condução deste mapeamento:

- *QP1: Quais são as linguagens de consultas mais usadas no modelo de dados em grafos ao longo dos anos?* Esta pergunta ajudará a identificar as linguagens mais usadas dentro do modelo de dados em grafos, se há novas linguagens sendo propostas e também

será possível analisar, de acordo com o número de publicações, se certas linguagem estão definindo tendências.

- *QP2: Quantas dessas linguagens possuem suporte a agregação?* As funções de agregação são muito úteis quando é necessário fazer algum tipo de cálculo ou medida que envolve o agrupamento do resultado de uma consulta. O modelo de dados relacional tem suporte para a agregação, então é importante verificar em números o que está sendo feito para o modelo de dados em grafos para suprir essa necessidade.

- *QP3: Quais paradigmas são mais usados na definição das linguagens de consulta para o modelo de dados em grafos?* Esta questão tem o objetivo de identificar as tendências na definição das linguagens, e também se isso vem mudando ao longo dos anos.

3.2 Condução da pesquisa

A primeira parte consistiu em definir a *string* de busca a partir das questões de pesquisa. As principais palavras-chave definidas foram: "grandes volumes de dados", "linguagem de consulta", "modelo de dados em grafo" e "funções de agregação". A busca deve recuperar artigos publicados nas principais bases de dados. Dessa forma, os termos devem ser buscados em inglês, haja vista que as principais publicações são feitas nesta língua. Assim sendo, as palavras-chave da busca usadas foram: *big data*, *query language*, *graph database* e *aggregation*. A partir destas palavras foram criadas variações destas para que a abrangência da consulta fosse maior.

As variações devem ser criadas porque para chegar à *string* de consulta final antes são feitos alguns testes. Normalmente as primeiras consultas não trazem resultados satisfatórios, ou a consulta é abrangente demais e retorna um número muito alto de resultados, ou a consulta é restrita demais e retorna um número muito baixo de resultados. As variações incluem adicionar o plural de alguns termos, e/ou adicionar sinônimos.

Baseado nas palavras-chave foi criada a seguinte consulta: *(("big data" OR "large data" OR "large dataset" OR "social network" OR "social networks") AND ("query language") AND ("graph database" OR "graph databases" OR "graph data" OR "large graph") AND ("aggregation" OR "summarisation" OR "aggregate"))*.

A segunda parte consistiu na escolha das bases de dados a serem consultadas. Foram escolhidas três bases de dados: *Science Direct*, *IEEE Xplore*, e *ACM digital library*. Estas bases foram escolhidas porque (i) elas são disponibilizadas na Internet, (ii) são populares para a área da Ciência da Computação, e (iii) agregam um grande número de conferências e periódicos científicos. Estas bases são disponibilizadas pela plataforma Portal de Periódicos

Tabela 1. Resultado da execução da consulta nas bibliotecas digitais.

Base de dados	Endereço eletrônico	Artigos recuperados
Science Direct	*www.sciencedirect.com*	67
IEEE Xplore	*ieeexplore.ieee.org*	89
ACM digital library	*dl.acm.org*	105
Total		**261**

CAPES[6], facilitando assim o acesso aos artigos. A plataforma Google Acadêmico[7] também foi considerada para ser incluída nas bases, mas ela foi descartada pelo fato de que grande parte dos resultados retornados já constarem nas outras bibliotecas digitais selecionadas, e ela também retorna muitos artigos de baixa representatividade.

A terceira parte desta etapa do processo consistiu na execução da *string* de busca em cada uma das bases de dados escolhidas. A execução da consulta resultou em um total de 261 artigos. A quantidade de artigos recuperados por base de dados pode ser visualizado na Tabela 1. Na tabela, a primeira coluna indica o nome da base de dados, seguido do endereço eletrônico desta base na Internet, e a quantidade de artigos recuperados.

3.3 Triagem dos artigos

Esta fase do mapeamento consiste em identificar quais artigos farão parte do mapeamento sistemático ou não, levando em conta os critérios estabelecidos. Os critérios de inclusão e exclusão de artigos foram definidos baseados no objetivo do mapeamento, levando em conta as questões de pesquisa. Os critérios definidos para este mapeamento podem ser lidos na Tabela 2.

Dentre os critérios de inclusão, os trabalhos devem mencionar alguma relação com linguagem de consulta e ter sido publicado em algum periódico, conferência ou *workshop*. Já os critérios de exclusão, cada um deles é um item eliminatório, ou seja, se um artigo se enquadra em algum dos critérios ele é eliminado.

Os trabalhos que não mencionam o uso de linguagens de consulta foram excluídos porque fugiam do escopo do mapeamento. Os capítulos de livros não foram considerados porque apresentam conteúdo que não é interessante para os números do mapeamento e sim para o uso no referencial teórico. A opção pela língua inglesa ocorreu para limitar os artigos que fossem possíveis de ser lidos pela autora, como também para eliminar possíveis artigos de mais baixa expressividade, escritos em língua nativa. Os anais completos abrangem uma

[6]http://www.periodicos.capes.gov.br/
[7]https://scholar.google.com.br/

Tabela 2. Critérios de inclusão e exclusão de artigos.

Inclusão
Trabalhos que mencionam linguagens de consulta em grafos nos resumos.
Artigos publicados em periódicos científicos, conferências ou *workshops*.
Exclusão
Trabalhos que não mencionam o uso de linguagens de consulta em grafos nos resumos.
Capítulos de livros.
Artigos escritos em língua diferente da língua inglesa.
Anais de congressos completos e índices.
Artigos repetidos.
Artigos sem resumo disponível.

grande quantidade de artigos com temas diversos que fogem ao escopo do mapeamento e os índices somente trazem os títulos, sem resumos. Os artigos sem resumo não entram no mapeamento, porque este mapeamento é feito baseado nos títulos e resumos. Um dos critérios de exclusão é a verificação de artigos repetidos, e esta verificação foi a primeira etapa da triagem.

Com os artigos organizados em listas, uma lista por base de dados, foi feita uma triagem em busca de trabalhos repetidos. Esta triagem foi feita com a leitura dos títulos dos artigos e a comparação entre as listas. Artigos com títulos iguais eram separados para uma verificação mais precisa do conteúdo. Estes eram verificados se apresentavam o(s) mesmo(s) autor(es) e o mesmo resumo. Ao constatar que o artigo era o mesmo, este era marcado como excluído, mantendo apenas um dos exemplares. Alguns artigos foram identificados como repetidos apenas após a leitura do resumo. Isso acontece porque há artigos em que os títulos são diferentes mas trata do mesmo trabalho. Muitas vezes o segundo artigo é uma versão estendida para um periódico ou uma continuação de um trabalho inicialmente já publicado. Em casos onde a diferença não é substancial o artigo é marcado como repetido. Ao todo foram identificados e excluídos 11 artigos repetidos.

Em seguida, a triagem continuou com base na leitura do título e dos resumos dos artigos e levando em conta os demais critérios de inclusão e exclusão. Após a finalização da triagem foram excluídos um total de 149 artigos. A totalização dos artigos pode ser vista na Tabela 3. A lista final contém um total de 101 artigos.

Tabela 3. Totalização após inclusão e exclusão de artigos.

	Science Direct	**IEEE**	**ACM**	**Total**
Artigos recuperados	67	89	105	261
Artigos repetidos	2	1	8	11
Artigos excluídos	44	53	52	149
Total	21	35	45	**101**

3.4 Termos-chave

A classificação dos artigos foi feita baseando-se em termos-chave que foram definidos, levando-se em conta as questões de pesquisa. A classificação levou à criação de três categorias. Cada categoria foi dividida em subcategorias para facilitar o processo de análise. As categorias e suas respectivas subcategorias são descritas a seguir.

Tipo de contribuição Esta categoria refere-se à principal contribuição do trabalho que está sendo analisado, o que está sendo proposto como objetivo principal. As divisões propostas para esta categoria foram (a) *linguagem* que indica se o trabalho apresenta uma nova linguagem ou a extensão de uma já existente; (b) *método* que indica se o trabalho apresenta um método para a resolução de um problema; (c) *survey/estado-da-arte/comparativo*; e ainda duas subcategorias complementares que indicam se o trabalho realizou (d) *experimentos* e disponibilizou (e) *ferramenta*.

Tipo de problema Refere-se ao tipo de problema que os autores pretendem abordar. Esta categoria apresenta as seguintes divisões: (a) *processamento de consultas* que indica se o foco da contribuição está em melhorias para o processamento de consultas; (b) *manipulação de grandes grafos* que indica se o trabalho está apto a lidar com grafos muito grandes (*big data*). (c) *funções de agregação* que indica se a linguagem implementa funções de agregação; (d) *consultas por palavras-chave* que indica o uso de palavras-chave para fazer consultas; (e) *top-k* que são aquelas linguagens que fazem consultas com um número limitado (preestabelecido) de resultados; (f) *integração de dados* quando o objetivo é unir dados de fontes diferentes; (g) *otimização de desempenho* que indica a implementação de alguma técnica de melhora no desempenho do algoritmo.

Paradigma tecnológico Refere-se ao modelo de programação, paradigma de programação ou ao tipo de tecnologia empregada a qual o problema envolve. Esta categoria apresenta as seguintes divisões: (a) *RDF* que é um modelo de dados muito usado para disponibilizar informações na *Web*; (b) *paralelismo* que indica o modelo paralelo de

programação; (c) *distribuição de dados* que indica manipulação de dados distribuídos; (d) *linguagem declarativa* que indica o uso do paradigma declarativo.

Os artigos incluídos no mapeamento foram classificados de acordo com as três categorias acima listadas. O resultado da classificação foi sumarizado nas Tabelas 4 a 6 que estão em um Anexo no final do artigo. As tabelas contém as referências dos artigos separados por categoria e subcategorias. Os dados foram analisados quantitativamente levando em conta a classificação proposta e enfatizando o número de publicações em cada categoria. A próxima seção mostrará a última etapa do processo de mapeamento sistemático, onde os resultados das análises feitas serão mostrados e as questões de pesquisa respondidas.

4 Análise e Discussões

Nesta seção serão apresentadas análises quantitativas e qualitativas do mapeamento sistemático relacionado às linguagens de consulta em grafos. A análise quantitativa apresentará os números de acordo com a classificação que foi indicada na seção anterior. A análise qualitativa apresentará relações entre os resultados quando for o caso, e discussões acerca dos números.

Recentemente, numerosos projetos para processamento, consulta e análise de bases de dados em grafos tem surgido [78]. Este retorno de interesse por esse modelo pode ser observado no gráfico da Figura 2. Este gráfico, lista a quantidade de artigos incluídos no mapeamento publicados por ano.

No gráfico é possível observar um crescimento do número de publicações a partir de 2008. Este mapeamento sistemático recuperou artigos publicados entre 1989 e 2015. Do total de artigos utilizados neste mapeamento 85% deles foram publicados entre 2009 e 2014. A condução da pesquisa de artigos nas bibliotecas foi feita até o dia 06/11/2014, por essa razão a quantidade de artigos publicados em 2015 está baixa. Antes de 2009 as publicações referentes às linguagens de consulta eram mais numerosas na área de XML[10]. Pode-se observar no gráfico que o número de artigos publicados nesta época na área de linguagens de consulta em grafos não foi expressivo. Antes da década de 2000 o número de publicações foi pequeno. Os anos de 1989, 1990, 1992, 1993, 1996 e 1999 tiveram um artigo em cada ano, e dois artigos no ano 2000.

O gráfico da Figura 3 apresenta o número de publicações por ano e por base de dados. O gráfico indica as três bases de dados usadas nesse mapeamento: SD (*Science Direct*), IEEE (*IEEE explore*) e ACM (*ACM digital library*). É possível observar que a base ACM se mantém com o maior número de publicações ao longo dos anos. Os artigos repetidos e que foram removidos não tem influência nesse resultado já que foi desta base a maior quantidade de artigos repetidos que foram excluídos. Isso pode ser constatado na Tabela 3.

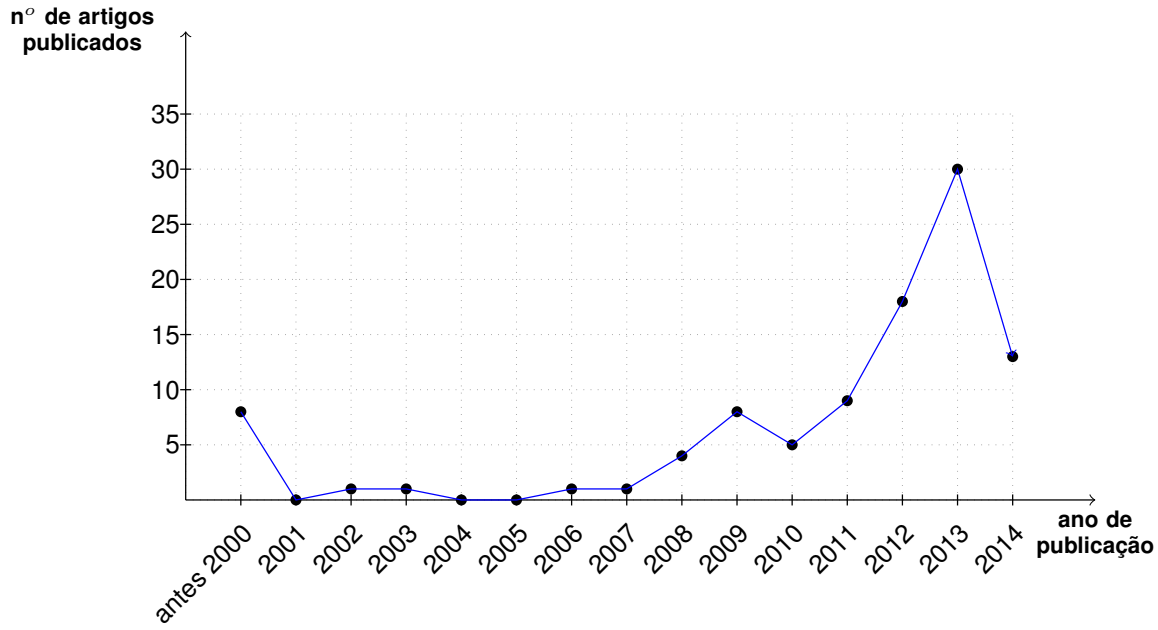


Figura 2. Publicações por ano.

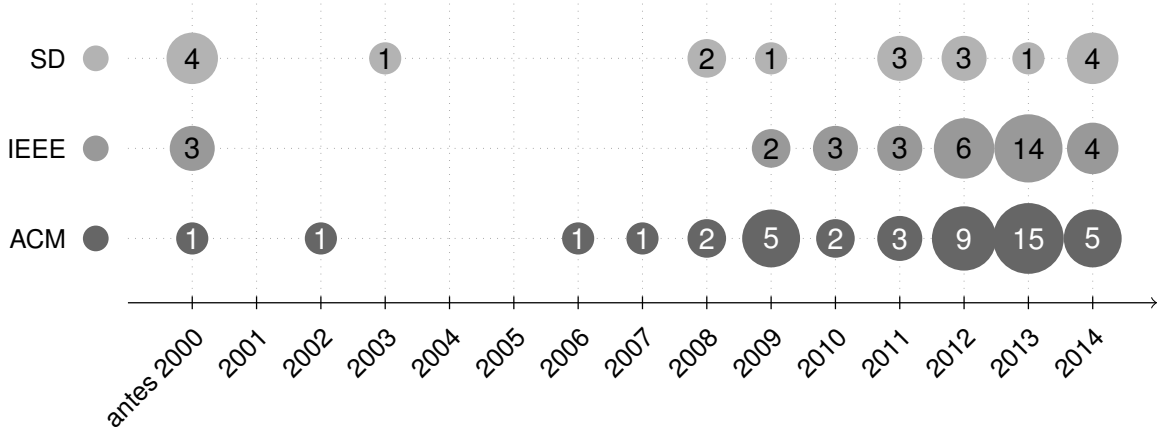


Figura 3. Publicações por ano e editora.

Há vários fatores que podem explicar esse retorno do modelo de dados em grafos. As redes sociais digitais, que ganharam popularidade nos últimos anos, são comumente representadas pelo modelo de dados em grafos. Os grafos são uma boa opção para as redes sociais por causa da sua característica de interconectividade, onde os vértices são entidades e as arestas são os relacionamentos entre as entidades. Nesses contextos a topologia, ou seja

a estrutura do grafo, é importante [3]. Neste mapeamento os tópicos relacionados às redes sociais são abordados nos resumos de 25% dos artigos selecionados.

Os grandes volumes de dados ou grandes grafos são gerados pelas redes sociais digitais e também estão presentes em diversos outros segmentos como análise social e *business intelligence* [40] [69] [34] [9] [90], farmacologia [37], partição e gerenciamento de dados [98] [87] [19] [81] [105], catálogos de referências bibliográficas [80], entre outros. Os tópicos relacionados a *big data* são abordados nos resumos de 48% dos artigos listados.

4.1 Análise das categorias

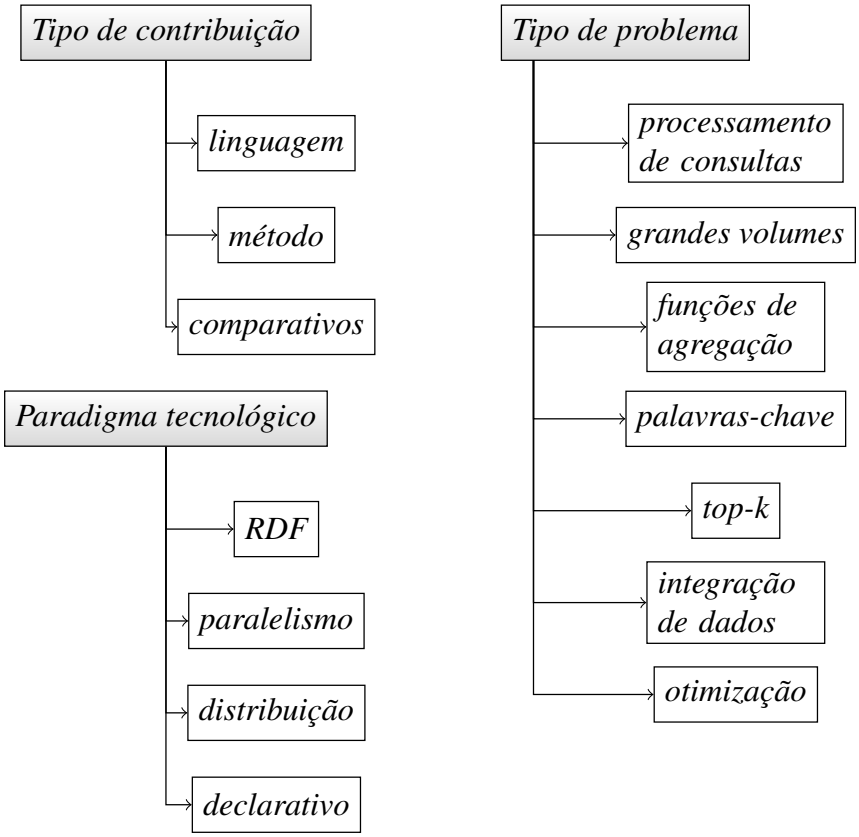


Figura 4. Divisão das categorias

Nesta seção serão analisadas as categorias de forma quantitativa, e também serão destacadas percepções a partir dos números apresentados. A Figura 4 apresenta um organograma de como as categorias foram divididas.

4.1.1 Tipo de contribuição Esta categoria teve como propósito indicar a principal contribuição do trabalho, o seu objetivo. A classificação do tipo de contribuição atribuía a um artigo se ele descrevia uma nova linguagem, um método ou conduzia um estudo comparativo. Como complemento a classificação também indicava se o artigo apresentava experimentos, e disponibilizava uma ferramenta.

Os artigos que apresentam uma nova linguagem de consulta ou uma extensão de uma linguagem já existente totalizaram 23%. Os artigos que apresentavam um método totalizaram 62%, e os estudos comparativos, que englobam *surveys* e estado-da-arte (visão atual sobre uma determinada área de estudo) totalizaram 17%. De todos esses artigos, 36% conduziram e apresentaram experimentos, e apenas 7% disponibilizavam ferramentas.

O número de artigos que conduziram e apresentaram experimentos foi considerada pequena pelos autores, sendo que menos da metade os apresentaram (36%). O experimentos são importantes na validação de um método ou ferramenta. Eles também são importantes para dar a possibilidade da audiência poder replicar o experimento e fazer comparativos.

4.1.2 Tipo de problema Esta categoria teve como propósito indicar o tipo de problema que o artigo pretende atacar. A classificação do tipo de problema se deu da seguinte forma: técnicas de processamento de consultas, gerenciamento de grandes volumes de dados, implementação de funções de agregação, técnicas de consultas por palavras-chave, técnicas de consultas *top-k*, integração de dados, e métodos de otimização de desempenho.

Os artigos que tem como foco técnicas de aperfeiçoamento do processamento de consultas são a maioria, representam 90%. Uma grande parte dos trabalhos está interessado em resolver problemas com o gerenciamento de grandes volumes de dados, estes representam 48%. A implementação de funções de agregação foi citada em 11% dos artigos. Consultas por palavras-chave e integração de dados representam 11% cada. Técnicas de consultas *top-k*, onde o número de resultados de uma consulta é limitado a um valor *k* representam 10%. E por fim, os artigos que implementam métodos de otimização de desempenho somaram 12%.

Como o foco principal deste mapeamento está nas linguagens de consulta para grafos era esperado que a maioria dos artigos tivessem como foco algum método ou algoritmo para melhoramento do processamento de consultas de uma forma geral. Dentro desta categoria, também foi possível observar que dentre todos os artigos que lidam com grandes volumes de dados (48% do total), 66% destes também implementam métodos para otimizar o seu desempenho. Lidar com um grande volume de dados frequentemente vem associado à busca de soluções para a melhoria do desempenho.

O aparecimento de diversos artigos que focam em consultas *top-k* mostra que nem sempre resultados completos são necessários. Dependendo do contexto, limitar estes resultados pode ser uma boa solução. Alguns trabalhos que lidam com *big data* como [71] e

[51] implementam consultas *top-k* que mostram apenas os resultados mais relevantes, seja porque os resultados completos não são necessários ou por uma questão de desempenho do algoritmo.

4.1.3 Paradigma tecnológico Esta categoria teve como propósito identificar o modelo de programação, paradigma de programação ou o tipo de tecnologia empregada. O paradigma de interesse para este mapeamento foi o declarativo, também entraram na classificação os grafos do tipo RDF, processamento paralelo e a distribuição de dados.

O paradigma declarativo foi detectado em 37% do total de artigos. Dentre os artigos que apresentam novas linguagens, 70% destas novas linguagens usam o paradigma declarativo. Os artigos que usam o modelo de dados RDF totalizaram 23%. Artigos que utilizam processamento paralelo somaram 12% e a distribuição de dados foi identificada em 8%. O uso do processamento paralelo, embora bastante popular, apresentou baixa percentagem neste mapeamento. Isso é explicado pelo fato de que o contexto deste trabalho está mais voltado para bancos de dados. O uso de termos como "linguagem de consulta" (*query language*) limita o aparecimento de trabalhos mais voltados para o processamento paralelo.

4.2 Relação entre as categorias

As Figuras 5 e 6 apresentam dois gráficos que fazem uma relação entre as categorias. O gráfico da Figura 5 apresenta uma relação entre a categoria "Tipo de contribuição" com as categorias "Tipo de problema" e "Paradigma tecnológico". O gráfico da Figura 6 apresenta uma relação entre as categorias "Tipo de problema" e "Paradigma tecnológico". Nestes gráficos é possível observar o número de publicações relacionando as subcategorias entre si.

Através do gráfico mostrado na Figura 5 é possível visualizar um maior volume de publicações relacionadas à melhoria no processamento de consultas e situações que envolvem grandes volumes de dados (lado esquerdo do gráfico). No lado direito do gráfico é possível visualizar que os grafos RDF também tem presença expressiva, especialmente com relação subcategoria "Método". Também é expressiva a relação entre as contribuições "Linguagem" e "Método" e o paradigma "Declarativo", onde 70% (16/23) das linguagem e 30% (18/62) dos métodos estão relacionados ao paradigma declarativo.

No gráfico da Figura 6 é possível fazer a seguinte observação: 61% (14/23) dos trabalhos que usam grafos RDF como base de dados estão associados com os grandes grafos. E ainda, (6/11) 55% implementam funções de agregação.

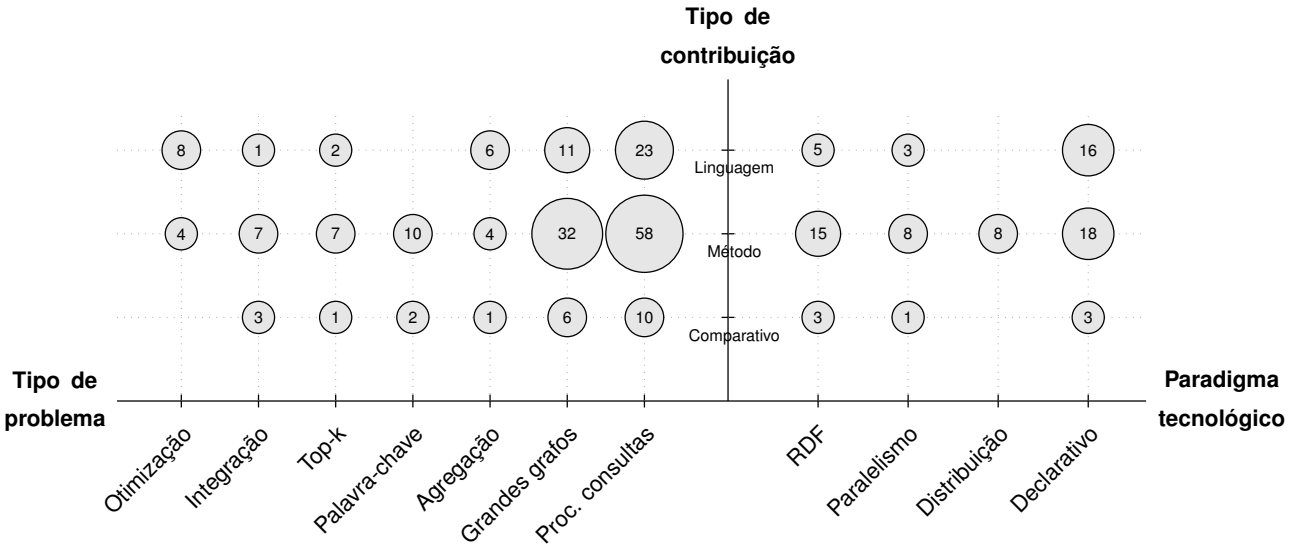


Figura 5. Relação entre a categoria "Tipo de contribuição" e as categorias "Tipo de problema" e "Paradigma tecnológico".

4.3 Questões de pesquisa

Nesta seção será apresentada uma análise baseada nas questões de pesquisa que foram lançadas no início deste mapeamento.

- *QP1: Quais são as linguagens de consulta mais usadas no modelo de dados em grafo ao longo dos anos?*

  Como não há uma linguagem padrão para uma base de dados em grafo, cada linguagem é criada ou estendida para atender ao seu problema específico [12]. A linguagem mais usada como base para a criação de novas linguagens é SQL (20% do total), a linguagem padrão para o modelo relacional, ou pelo menos estas novas linguagens usam uma sintaxe muito similar. Dentre elas podemos citar SQL/TC [23], BiQL [30], SoQL [79], SCOPE [108], e o trabalho apresentado em [97]. Dentre as baseadas em Datalog, uma linguagem de consulta para base de dados dedutivos, podemos citar SociaLite [84], DatalogFS [65], Glog [34]. Dentre as baseadas em SPARQL, a linguagem de consulta recomendada para RDF, tem-se G-SPARQL [80], MashQL [46], SPARankQL [15]. A linguagem Cypher, que é a linguagem de consulta para sistemas Neo4j, foi usada como referência em [95][60][14]. Cypher também tem sido usada em comparativos, como em [42][4][43]. Outras linguagens que serviram como referência para novas linguagens são OQL e XPath.

- *QP2: Quantas dessas linguagens possuem suporte à agregação?*

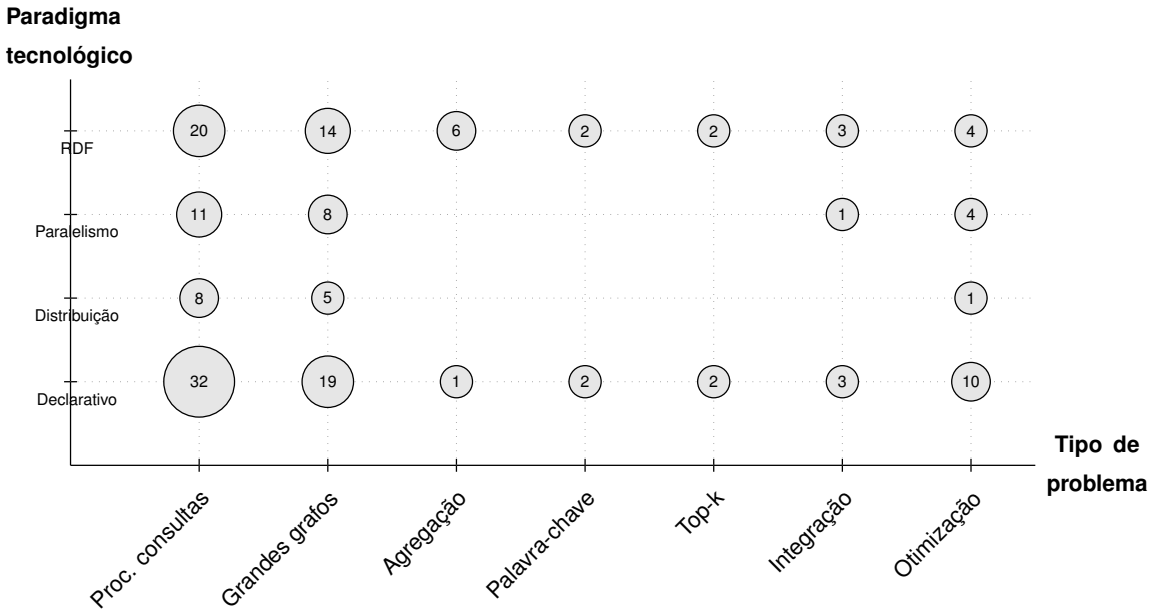


Figura 6. Relação entre "Tipo de problema" e "Paradigma tecnológico".

A agregação é uma função bastante usada no modelo relacional, mas que não tem a mesma popularidade no modelo de dados para grafo. Uma função de agregação usa o resultado de uma consulta como valor intermediário para fazer uma outra operação. As funções de agregação mais comuns são soma, média, maior valor e menor valor. Os trabalhos que apontaram para o uso de funções de agregação somam apenas 11% do total. Aqui pode-se sinalizar uma área com potencial a ser explorado.

- *QP3: Quais paradigmas são mais usados na definição das linguagens de consulta para o modelo de dados em grafos?*

  Das linguagens que foram identificadas neste mapeamento, a maioria delas, 70% são linguagens declarativas. Dentre estas, a maior parte usam o estilo de consultas "`SELECT -WHERE -FROM`" que é conhecido como o estilo SQL. O uso desse estilo é justificado muitas vezes pelo fato de ser muito conhecido, o que tornaria a compreensão da nova linguagem mais rápida.

5 Conclusões

Este mapeamento sistemático apresentou uma investigação acerca de linguagens de consulta para base de dados em grafos. Foram usados um total de 101 artigos que foram

distribuídos em categorias que facilitaram a extração e análise de informações sobre o tema.

A partir desta análise foi concluído que o uso da base de dados em grafos teve um aumento no número de publicações a partir da década de 2000. Com o avanço da internet e das redes sociais, a busca por diferentes tecnologias além do modelo relacional ficou evidenciada.

O modelo de dados em grafos não possui uma linguagem de consulta padrão, a maioria das contribuições são específicas para o domínio em que é aplicado. No entanto, os grafos RDF que possuem uma recomendação (*W3C recomendation*) para linguagem de consulta, usam o SPARQL como referência para novas linguagens e extensões.

Os números deste mapeamento destacaram o uso do paradigma declarativo para a construção de novas linguagens, sem indicativo claro de uma mudança nesse sentido para os próximos anos. Também foi identificado que as funções de agregação não são muito exploradas nas linguagens de consulta, indicando que este tópico tem potencial para trabalhos futuros.

Além das conclusões específicas deste estudo, o mesmo serviu para repertoriar as principais contribuições da área de linguagens de consulta para bases de dados em grafos, servindo, assim, como fonte de referências bibliográficas da área, em complemento a outras fontes como [101][3][5].

Contribuição dos autores:

- Simone de Oliveira SANTOS: Conduziu a pesquisa e escreveu o artigo.

- Mirian Halfeld Ferrari Alves e Martin A. Musicante: Supervisão da condução da pesquisa. Os co-autores indicaram bases de dados, trabalharam nas formulação das strings de busca e revisaram a escrita do artigo. Também foram os responsáveis pelo direcionamento referente aos tipos de gráficos apresentados no trabalho e associações a serem feitas.

Referências

[1] R. Angles. A comparison of current graph database models. In *IEEE 28th International Conference on Data Engineering Workshops (ICDEW)*, pages 171–177, April 2012.

[2] Renzo Angles and Claudio Gutierrez. Querying RDF data from a graph database perspective. In Asunción Gómez-Pérez and Jérôme Euzenat, editors, *The Semantic Web: Research and Applications*, volume 3532 of *Lecture Notes in Computer Science*, pages 346–360. Springer Berlin Heidelberg, 2005.

[3] Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Comput. Surv.*, 40(1):1:1–1:39, February 2008.

[4] Renzo Angles, Arnau Prat-Pérez, David Dominguez-Sal, and Josep-Lluis Larriba-Pey. Benchmarking database systems for social network applications. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 15:1–15:7, New York, NY, USA, 2013. ACM.

[5] James Bailey, François Bry, Tim Furche, and Sebastian Schaffert. Web and semantic web languages: A survey. In Norbert Eisinger and Jan Maluszynski, editors, *Reasoning Web*, volume 3564 of *Lecture Notes in Computer Science*, pages 35–133. Springer Berlin Heidelberg, 2005.

[6] Pablo Barceló Baeza. Querying graph databases. In *Proceedings of the 32nd Symposium on Principles of Database Systems*, PODS '13, pages 175–188, New York, NY, USA, 2013. ACM.

[7] Yijun Bei, Zhen Lin, Chen Zhao, and Xiaojun Zhu. HBase system-based distributed framework for searching large graph databases. In *14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 151–156, July 2013.

[8] Tim Berners-Lee and Jeffrey Jaffe. World Wide Web Consortium (W3C).

[9] Dritan Bleco and Yannis Kotidis. Business intelligence on complex graph data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, pages 13–20, New York, NY, USA, 2012. ACM.

[10] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau, and John Cowan. Extensible Markup Language (XML) 1.1. W3C Recomendation. August 2006.

[11] M. Brocheler, A. Pugliese, and V.S. Subrahmanian. COSI: Cloud oriented subgraph identification in massive social networks. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 248–255, Aug 2010.

[12] Mike Buerli and Cal Poly San Luis Obispo. The current state of graph databases. *Department of Computer Science, Cal Poly San Luis Obispo, mbuerli@calpoly.edu*, 2012.

[13] V.G. Castellana, A. Tumeo, O. Villa, D. Haglin, and J. Feo. Composing data parallel code for a SPARQL graph engine. In *International Conference on Social Computing (SocialCom)*, pages 691–699, Sept 2013.

[14] Ciro Cattuto, Marco Quaggiotto, André Panisson, and Alex Averbuch. Time-varying social networks in a graph database: A Neo4J use case. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 11:1–11:6, New York, NY, USA, 2013. ACM.

[15] Juan P. Cedeño and K. Selçuk Candan. R2DF framework for ranked path queries over weighted RDF graphs. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 40:1–40:12, New York, NY, USA, 2011. ACM.

[16] Alan Chappell, Sutanay Choudhury, John Feo, David Haglin, Alessandro Morari, Sumit Purohit, Karen Schuchardt, Antonino Tumeo, Jesse Weaver, and Oreste Villa. Toward a data scalable solution for facilitating discovery of scientific data resources. In *Proceedings of the 2013 International Workshop on Data-Intensive Scalable Computing Systems*, DISCS-2013, pages 55–60, New York, NY, USA, 2013. ACM.

[17] Yi Chen, Wei Wang, and Ziyang Liu. Keyword-based search and exploration on databases. In *IEEE 27th International Conference on Data Engineering (ICDE)*, pages 1380–1383, April 2011.

[18] M. P. Consens and A. O. Mendelzon. Expressing structural hypertext queries in graphlog. In *Proceedings of the Second Annual ACM Conference on Hypertext*, HYPERTEXT '89, pages 269–292, New York, NY, USA, 1989. ACM.

[19] O. Cure, F. Kerdjoudj, Chan Le Duc, M. Lamolle, and D. Faye. On the potential integration of an ontology-based data access approach in NoSQL stores. In *Third International Conference on Emerging Intelligent Data and Web Technologies (EIDWT)*, pages 166–173, Sept 2012.

[20] Michael Curtiss, Iain Becker, Tudor Bosman, Sergey Doroshenko, Lucian Grijincu, Tom Jackson, Sandhya Kunnatur, Soren Lassen, Philip Pronin, Sriram Sankar, Guanghao Shen, Gintaras Woss, Chao Yang, and Ning Zhang. Unicorn: A system for searching the social graph. *Proc. VLDB Endow.*, 6(11):1150–1161, August 2013.

[21] A. Cuzzocrea and P. Serafino. A reachability-based theoretical framework for modeling and querying complex probabilistic graph data. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1177–1184, Oct 2012.

[22] Richard Cyganiak, David Wood, Markus Lanthaler, Graham Klyne, Jeremy J. Carroll, and Brian McBride. Resource Description Framework (RDF): Concepts and abstract syntax. W3C Recomendation. February 2014.

[23] S. Dar and R. Agrawal. Extending SQL with generalized transitive closure. *IEEE Transactions on Knowledge and Data Engineering*, 5(5):799–812, Oct 1993.

[24] Roberto De Virgilio, Antonio Maccioni, and Riccardo Torlone. Converting relational to graph databases. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 1:1–1:6, New York, NY, USA, 2013. ACM.

[25] Renaud Delbru, Stephane Campinas, and Giovanni Tummarello. Searching web data: An entity retrieval and high-performance indexing model. *Web Semantics: Science, Services and Agents on the World Wide Web*, 10(0):33 – 58, 2012. Web-Scale Semantic Information Processing.

[26] Saumen Dey, Víctor Cuevas-Vicenttín, Sven Köhler, Eric Gribkoff, Michael Wang, and Bertram Ludäscher. On implementing provenance-aware regular path queries with relational query engines. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 214–223, New York, NY, USA, 2013. ACM.

[27] Ying Ding. Community detection: Topological vs. topical. *Journal of Informetrics*, 5(4):498 – 514, 2011.

[28] Yerach Doytsher, Ben Galon, and Yaron Kanza. Querying geo-social data by bridging spatial networks and social networks. In *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, LBSN '10, pages 39–46, New York, NY, USA, 2010. ACM.

[29] Anton Dries and Siegfried Nijssen. Analyzing graph databases by aggregate queries. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, MLG '10, pages 37–45, New York, NY, USA, 2010. ACM.

[30] Anton Dries, Siegfried Nijssen, and Luc De Raedt. A query language for analyzing networks. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 485–494, New York, NY, USA, 2009. ACM.

[31] V. Dritsou, P. Constantopoulos, A. Deligiannakis, and Y. Kotidis. Shortcut selection in RDF databases. In *IEEE 27th International Conference on Data Engineering Workshops (ICDEW)*, pages 194–199, April 2011.

[32] Wenfei Fan, Xin Wang, and Yinghui Wu. Incremental graph pattern matching. *ACM Trans. Database Syst.*, 38(3):18:1–18:47, September 2013.

[33] L. Fegaras. Supporting Bulk Synchronous Parallelism in Map-Reduce queries. In *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*, pages 1068–1077, Nov 2012.

[34] Jun Gao, Jiashuai Zhou, Chang Zhou, and J.X. Yu. GLog: A high level graph analysis system using mapreduce. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 544–555, March 2014.

[35] A. González-Beltrán, P. Milligan, and P. Sage. Range queries over skip tree graphs. *Computer Communications*, 31(2):358 – 374, 2008. Special Issue: Foundation of Peer-to-Peer Computing.

[36] Stewart Greenhill and Svetha Venkatesh. Semantic data modelling and visualisation using noetica. *Data & Knowledge Engineering*, 33(3):241 – 276, 2000.

[37] Paul Groth, Antonis Loizou, Alasdair J.G. Gray, Carole Goble, Lee Harland, and Steve Pettifer. Api-centric linked data integration: The open PHACTS discovery platform case study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2014.

[38] Amarnath Gupta and Simone Santini. On querying OBO ontologies using a DAG pattern query language. In *Proceedings of the Third International Conference on Data Integration in the Life Sciences*, DILS'06, pages 152–167, Berlin, Heidelberg, 2006. Springer-Verlag.

[39] Hao He, Haixun Wang, Jun Yang, and Philip S. Yu. BLINKS: Ranked keyword searches on graphs. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 305–316, New York, NY, USA, 2007. ACM.

[40] J. Heer and A. Perer. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 51–60, Oct 2011.

[41] Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker. Searching and browsing linked data with SWSE: The semantic web search engine. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):365 – 401, 2011. {JWS} special issue on Semantic Search.

[42] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of linked data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14(0):14 – 44, 2012. Special Issue on Dealing with the Messiness of the Web of Data.

[43] Florian Holzschuher and René Peinl. Performance of graph query languages: Comparison of cypher, gremlin and native access in Neo4J. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 195–204, New York, NY, USA, 2013. ACM.

[44] T. Härder, B. Mitschang, and H. Schöning. Query processing for complex objects. *Data & Knowledge Engineering*, 7(3):181 – 200, 1992.

[45] H.M. Jamil. Design of declarative graph query languages: On the choice between value, pattern and object based representations for graphs. In *IEEE 28th International Conference on Data Engineering Workshops (ICDEW)*, pages 178–185, April 2012.

[46] M. Jarrar and M.D. Dikaiakos. Querying the data web: The MashQL approach. *Internet Computing, IEEE*, 14(3):58–67, May 2010.

[47] Ruoming Jin, Yang Xiang, Ning Ruan, and David Fuhry. 3-HOP: A high-compression indexing scheme for reachability query. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, pages 813–826, New York, NY, USA, 2009. ACM.

[48] S. Jouili and V. Vansteenberghe. An empirical comparison of graph databases. In *International Conference on Social Computing (SocialCom)*, pages 708–715, Sept 2013.

[49] Andrey Kan, Jeffrey Chan, James Bailey, and Christopher Leckie. A query based approach for mining evolving graphs. In *Proceedings of the Eighth Australasian Data Mining Conference - Volume 101*, AusDM '09, pages 139–150, Darlinghurst, Australia, Australia, 2009. Australian Computer Society, Inc.

[50] G. Karvounarakis, A. Magganaraki, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, and K. Tolle. Querying the semantic web with RQL. *Computer Networks*, 42(5):617 – 640, 2003. The Semantic Web: an evolution for a revolution.

[51] G. Kasneci, M. Ramanath, M. Sozio, F.M. Suchanek, and G. Weikum. STAR: Steiner-tree approximation in relationship graphs. In *IEEE 25th International Conference on Data Engineering - ICDE'09*, pages 868–879, March 2009.

[52] K. Kaur and R. Rani. Modeling and querying data in NoSQL databases. In *IEEE International Conference on Big Data*, pages 1–7, Oct 2013.

[53] A. Khan, Yinghui Wu, and X. Yan. Emerging graph queries in linked data. In *IEEE 28th International Conference on Data Engineering (ICDE)*, pages 1218–1221, April 2012.

[54] Arijit Khan, Nan Li, Xifeng Yan, Ziyu Guan, Supriyo Chakraborty, and Shu Tao. Neighborhood based fast graph search in large networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 901–912, New York, NY, USA, 2011. ACM.

[55] Benny Kimelfeld and Yehoshua Sagiv. Efficiently enumerating results of keyword search over data graphs. *Information Systems*, 33(45):335 – 359, 2008. Selected Papers from the Tenth International Symposium on Database Programming Languages (DBPL 2005.

[56] Taekyong Lee, Lei Sheng, T. Bozkaya, N.H. Balkir, Z. Meral Ozsoyoglu, and Z.M. Ozsoyoglu. Querying multimedia presentations based on content. *IEEE Transactions on Knowledge and Data Engineering*, 11(3):361–385, May 1999.

[57] M. Leida and A. Chu. Distributed SPARQL query answering over RDF data streams. In *IEEE International Congress on Big Data (BigData Congress)*, pages 369–378, June 2013.

[58] Xiang Lian, Eugenio De Hoyos, Artem Chebotko, Bin Fu, and Christine Reilly. k-nearest keyword search in RDF graphs. *Web Semantics: Science, Services and Agents on the World Wide Web*, pages 40 – 56, 2013.

[59] Vitaliy Liptchinsky, Benjamin Satzger, Rostyslav Zabolotnyi, and Schahram Dustdar. Expressive languages for selecting groups from graph-structured data. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 761–770, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[60] Yunkai Liu and T.M. Vitolo. Graph data warehouse: Steps to integrating graph databases into the traditional conceptual structure of a data warehouse. In *IEEE International Congress on Big Data (BigData Congress)*, pages 433–434, June 2013.

[61] Michel Mainguenaud. Constraint-based queries in a geographical database for network facilities. *Computers, Environment and Urban Systems*, 20(2):139 – 151, 1996.

[62] Federica Mandreoli, Riccardo Martoglia, Giorgio Villani, and Wilma Penzo. Flexible query answering on graph-modeled data. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 216–227, New York, NY, USA, 2009. ACM.

[63] Norbert Martinez-Bazan and David Dominguez-Sal. Using semijoin programs to solve traversal queries in graph databases. In *Proceedings of Workshop on GRAph Data Management Experiences and Systems*, GRADES'14, pages 6:1–6:6, New York, NY, USA, 2014. ACM.

[64] Yosi Mass and Yehoshua Sagiv. Language models for keyword search over data graphs. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 363–372, New York, NY, USA, 2012. ACM.

[65] Mirjana Mazuran, Edoardo Serra, and Carlo Zaniolo. Extending the power of datalog recursion. *The VLDB Journal*, 22(4):471–493, August 2013.

[66] Jayanta Mondal and Amol Deshpande. EAGr: Supporting continuous ego-centric aggregate queries over large dynamic graphs. In *Proceedings of the 2014 ACM SIGMOD*

International Conference on Management of Data, SIGMOD '14, pages 1335–1346, New York, NY, USA, 2014. ACM.

[67] A. Morari, V.G. Castellana, D. Haglin, J. Feo, J. Weaver, A. Tumeo, and O. Villa. Accelerating semantic graph databases on commodity clusters. In *IEEE International Conference on Big Data*, pages 768–772, Oct 2013.

[68] W.E. Moustafa, A. Kimmig, A. Deshpande, and L. Getoor. Subgraph pattern matching over uncertain graphs with identity linkage uncertainty. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 904–915, March 2014.

[69] A. Naseer, L. Laera, and T. Matsutsuka. Enterprise biggraph. In *46th Hawaii International Conference on System Sciences (HICSS)*, pages 1005–1014, Jan 2013.

[70] María Constanza Pabón, Claudia Roncancio, and Martha Millán. Graph data transformations and querying. In *Proceedings of the 2014 International C* Conference on Computer Science & Software Engineering*, C3S2E '14, pages 20:1–20:6, New York, NY, USA, 2014. ACM.

[71] Chang-Sup Park and Sungchae Lim. Efficient processing of keyword queries over graph databases for finding effective answers. *Information Processing & Management*, 51(1):42 – 57, 2015.

[72] Raqueline RM Penteado, Rebeca Schroeder, Diego Hoss, Jaqueline Nande, Ricardo M Maeda, Walmir O Couto, and Carmem S Hara. Um estudo sobre bancos de dados em grafos nativos. X ERBD - Escola Regional de Banco de Dados.

[73] A. Petermann, M. Junghanns, R. Muller, and E. Rahm. BIIIG: Enabling business intelligence with integrated instance graphs. In *IEEE 30th International Conference on Data Engineering Workshops (ICDEW)*, pages 4–11, March 2014.

[74] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic mapping studies in software engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, EASE'08, pages 68–77, Swinton, UK, 2008. British Computer Society.

[75] Letao Qi, H.T. Lin, and V. Honavar. Clustering remote RDF data using SPARQL update queries. In *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 236–242, April 2013.

[76] C. Raguenaud, J. Kennedy, and P.J. Barclay. The Prometheus taxonomic database. In *Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pages 63–70, 2000.

[77] Padmashree Ravindra, HyeongSik Kim, and Kemafor Anyanwu. To nest or not to nest, when and how much: Representing intermediate results of graph pattern queries in mapreduce based processing. In *Proceedings of the 4th International Workshop on Semantic Web Information Management*, SWIM '12, pages 5:1–5:8, New York, NY, USA, 2012. ACM.

[78] Ian Robinson, Jim Webber, and Emil Eifrem. *Graph Databases*. O'Reilly Media, Inc., 2013.

[79] R. Ronen and O. Shmueli. SoQL: A language for querying and creating data in social networks. In *ICDE'09. IEEE 25th International Conference on Data Engineering*, pages 1595–1602, March 2009.

[80] Sherif Sakr, Sameh Elnikety, and Yuxiong He. Hybrid query execution engine for large attributed graphs. *Information Systems*, 41(0):45 – 73, 2014.

[81] Semih Salihoglu and Jennifer Widom. GPS: A graph processing system. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, SSDBM, pages 22:1–22:12, New York, NY, USA, 2013. ACM.

[82] Mohamed Sarwat, Sameh Elnikety, Yuxiong He, and Mohamed F. Mokbel. Horton+: A distributed system for processing declarative reachability queries over partitioned graphs. *Proc. VLDB Endow.*, 6(14):1918–1929, September 2013.

[83] Timos K. Sellis, Nick Roussopoulos, and Raymond T. Ng. Efficient compilation of large rule bases using logical access paths. *Information Systems*, 15(1):73 – 84, 1990. Knowledge Engineering.

[84] Jiwon Seo, S. Guo, and M.S. Lam. SociaLite: Datalog extensions for efficient social network analysis. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 278–289, April 2013.

[85] Hong-Han Shuai, De-Nian Yang, P.S. Yu, Chih-Ya Shen, and Ming-Syan Chen. On pattern preserving graph generation. In *IEEE 13th International Conference on Data Mining (ICDM)*, pages 677–686, Dec 2013.

[86] Steve Speicher, John Arwe, and Ashok Malhotra. Linked data platform 1.0. W3C Recomendation. February 2015.

[87] S.R. Spillane, J. Birnbaum, D. Bokser, D. Kemp, A. Labouseur, P.W. Olsen, J. Vijayan, Jeong-Hyon Hwang, and Jun-Weon Yoon. A demonstration of the G* graph database system. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 1356–1359, April 2013.

[88] Dan Suciu. Distributed query evaluation on semistructured data. *ACM Trans. Database Syst.*, 27(1):1–62, March 2002.

[89] Zhao Sun, Hongzhi Wang, Haixun Wang, Bin Shao, and Jianzhong Li. Efficient subgraph matching on billion node graphs. *Proc. VLDB Endow.*, 5(9):788–799, May 2012.

[90] Yuanyuan Tian, Jignesh M. Patel, Viji Nair, Sebastian Martini, and Matthias Kretzler. Periscope/GQ: A graph querying toolkit. *Proc. VLDB Endow.*, 1(2):1404–1407, August 2008.

[91] Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 125–134, New York, NY, USA, 2012. ACM.

[92] Thanh Tran, Haofen Wang, and Peter Haase. Hermes: Data web search on a pay-as-you-go integration infrastructure. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):189 – 203, 2009. The Web of Data.

[93] Ying Tu and Han-Wei Shen. GraphCharter: Combining browsing with query to explore large semantic graphs. In *Visualization Symposium (PacificVis), 2013 IEEE Pacific*, pages 49–56, Feb 2013.

[94] M. Uddin, R. Stadler, M. Miyazawa, and M. Hayashi. Graph search for cloud network management. In *Network Operations and Management Symposium (NOMS), 2014 IEEE*, pages 1–5, May 2014.

[95] Raoul-Gabriel Urma and Alan Mycroft. Source-code queries with graph databases with application to programming language usage and evolution. *Science of Computer Programming*, 97, Part 1(0):127 – 134, 2015. Special Issue on New Ideas and Emerging Results in Understanding Software.

[96] Ramakrishna Varadarajan, Vagelis Hristidis, Louiqa Raschid, Maria-Esther Vidal, Luis Ibáñez, and Héctor Rodríguez-Drumond. Flexible and efficient querying and ranking on hyperlinked data sources. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 553–564, New York, NY, USA, 2009. ACM.

[97] Fan Wang and Gagan Agrawal. Answering complex structured queries over the deep web. In *Proceedings of the 15th Symposium on International Database Engineering & Applications*, IDEAS '11, pages 115–123, New York, NY, USA, 2011. ACM.

[98] Rui Wang and K. Chiu. A stream partitioning approach to processing large scale distributed graph datasets. In *IEEE International Conference on Big Data*, pages 537–542, Oct 2013.

[99] Jesse Weaver, Vito Giovanni Castellana, Alessandro Morari, Antonino Tumeo, Sumit Purohit, Alan Chappell, David Haglin, Oreste Villa, Sutanay Choudhury, Karen Schuchardt, and John Feo. Toward a data scalable solution for facilitating discovery of science resources. *Parallel Computing*, 40(10):682 – 696, 2014.

[100] Klara Weiand, Fabian KneiSSl, Wojciech Lobacz, Tim Furche, and François Bry. pest: Fast approximate keyword search in semantic data using eigenvector-based term propagation. *Information Systems*, 37(4):372 – 390, 2012. Semantic Web Data Management.

[101] Peter T. Wood. Query languages for graph databases. *SIGMOD Rec.*, 41(1):50–60, April 2012.

[102] Wenlei Xie, Guozhang Wang, David Bindel, Alan Demers, and Johannes Gehrke. Fast iterative graph computation with block updates. *Proc. VLDB Endow.*, 6(14):2014–2025, September 2013.

[103] P. Yadav and V. Samala. Benchmarking over a semantic repository. In *Second International Conference on Advanced Computing (ICoAC)*, pages 51–59, Dec 2010.

[104] Shengqi Yang, Yanan Xie, Yinghui Wu, Tianyu Wu, Huan Sun, Jian Wu, and Xifeng Yan. SLQ: A user-friendly graph querying system. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 893–896, New York, NY, USA, 2014. ACM.

[105] Shengqi Yang, Xifeng Yan, Bo Zong, and Arijit Khan. Towards effective partition management for large graphs. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 517–528, New York, NY, USA, 2012. ACM.

[106] Hilmi Yildirim, Vineet Chaoji, and Mohammed J. Zaki. GRAIL: A scalable index for reachability queries in very large graphs. *The VLDB Journal*, 21(4):509–534, August 2012.

[107] Kai Zeng, Jiacheng Yang, Haixun Wang, Bin Shao, and Zhongyuan Wang. A distributed graph engine for web scale RDF data. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pages 265–276. VLDB Endowment, 2013.

[108] Jingren Zhou, Nicolas Bruno, Ming-Chuan Wu, Per-Ake Larson, Ronnie Chaiken, and Darren Shakib. SCOPE: Parallel databases meet MapReduce. *The VLDB Journal*, 21(5):611–636, October 2012.

[109] Linhong Zhu, Wee Keong Ng, and James Cheng. Structure and attribute index for approximate graph matching in large graphs. *Information Systems*, 36(6):958 – 972, 2011.

[110] Lei Zou, M. Tamer Özsu, Lei Chen, Xuchuan Shen, Ruizhe Huang, and Dongyan Zhao. gStore: A graph-based SPARQL query engine. *The VLDB Journal*, 23(4):565–590, August 2014.

ANEXO

Tabela 4: Categoria: tipo de contribuição.

Referências dos artigos da subcategoria Linguagem
- Sherif Sakr, Sameh Elnikety, and Yuxiong He. Hybrid query execution engine for large attributed graphs. *Information Systems*, 41(0):45 – 73, 2014
- G. Karvounarakis, A. Magganaraki, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, and K. Tolle. Querying the semantic web with RQL. *Computer Networks*, 42(5):617 – 640, 2003. The Semantic Web: an evolution for a revolution
- R. Ronen and O. Shmueli. SoQL: A language for querying and creating data in social networks. In *ICDE'09. IEEE 25th International Conference on Data Engineering*, pages 1595–1602, March 2009
- J. Heer and A. Perer. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 51–60, Oct 2011
- Jiwon Seo, S. Guo, and M.S. Lam. SociaLite: Datalog extensions for efficient social network analysis. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 278–289, April 2013
- C. Raguenaud, J. Kennedy, and P.J. Barclay. The Prometheus taxonomic database. In *Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pages 63–70, 2000
- Jun Gao, Jiashuai Zhou, Chang Zhou, and J.X. Yu. GLog: A high level graph analysis system using mapreduce. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 544–555, March 2014
- P. Yadav and V. Samala. Benchmarking over a semantic repository. In *Second International Conference on Advanced Computing (ICoAC)*, pages 51–59, Dec 2010
- Taekyong Lee, Lei Sheng, T. Bozkaya, N.H. Balkir, Z. Meral Ozsoyoglu, and Z.M. Ozsoyoglu. Querying multimedia presentations based on content. *IEEE Transactions on Knowledge and Data Engineering*, 11(3):361–385, May 1999
- S. Dar and R. Agrawal. Extending SQL with generalized transitive closure. *IEEE Transactions on Knowledge and Data Engineering*, 5(5):799–812, Oct 1993
- M. Jarrar and M.D. Dikaiakos. Querying the data web: The MashQL approach. *Internet Computing, IEEE*, 14(3):58–67, May 2010
- Anton Dries, Siegfried Nijssen, and Luc De Raedt. A query language for analyzing networks. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 485–494, New York, NY, USA, 2009. ACM
- Anton Dries and Siegfried Nijssen. Analyzing graph databases by aggregate queries. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, MLG '10, pages 37–45, New York, NY, USA, 2010. ACM
- Fan Wang and Gagan Agrawal. Answering complex structured queries over the deep web. In *Proceedings of the 15th Symposium on International Database Engineering & Applications*, IDEAS '11, pages 115–123, New York, NY, USA, 2011. ACM
- M. P. Consens and A. O. Mendelzon. Expressing structural hypertext queries in graphlog. In *Proceedings of the Second Annual ACM Conference on Hypertext*, HYPERTEXT '89, pages 269–292, New York, NY, USA, 1989. ACM

- Vitaliy Liptchinsky, Benjamin Satzger, Rostyslav Zabolotnyi, and Schahram Dustdar. Expressive languages for selecting groups from graph-structured data. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 761–770, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee
- Mirjana Mazuran, Edoardo Serra, and Carlo Zaniolo. Extending the power of datalog recursion. *The VLDB Journal*, 22(4):471–493, August 2013
- Mohamed Sarwat, Sameh Elnikety, Yuxiong He, and Mohamed F. Mokbel. Horton+: A distributed system for processing declarative reachability queries over partitioned graphs. *Proc. VLDB Endow.*, 6(14):1918–1929, September 2013
- Amarnath Gupta and Simone Santini. On querying OBO ontologies using a DAG pattern query language. In *Proceedings of the Third International Conference on Data Integration in the Life Sciences*, DILS'06, pages 152–167, Berlin, Heidelberg, 2006. Springer-Verlag
- Yerach Doytsher, Ben Galon, and Yaron Kanza. Querying geo-social data by bridging spatial networks and social networks. In *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, LBSN '10, pages 39–46, New York, NY, USA, 2010. ACM
- Juan P. Cedeño and K. Selçuk Candan. R2DF framework for ranked path queries over weighted RDF graphs. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 40:1–40:12, New York, NY, USA, 2011. ACM
- Jingren Zhou, Nicolas Bruno, Ming-Chuan Wu, Per-Ake Larson, Ronnie Chaiken, and Darren Shakib. SCOPE: Parallel databases meet MapReduce. *The VLDB Journal*, 21(5):611–636, October 2012
- Michael Curtiss, Iain Becker, Tudor Bosman, Sergey Doroshenko, Lucian Grijincu, Tom Jackson, Sandhya Kunnatur, Soren Lassen, Philip Pronin, Sriram Sankar, Guanghao Shen, Gintaras Woss, Chao Yang, and Ning Zhang. Unicorn: A system for searching the social graph. *Proc. VLDB Endow.*, 6(11):1150–1161, August 2013

.

Referências dos artigos subcategoria Método

- Linhong Zhu, Wee Keong Ng, and James Cheng. Structure and attribute index for approximate graph matching in large graphs. *Information Systems*, 36(6):958 – 972, 2011
- Chang-Sup Park and Sungchae Lim. Efficient processing of keyword queries over graph databases for finding effective answers. *Information Processing & Management*, 51(1):42 – 57, 2015
- Renaud Delbru, Stephane Campinas, and Giovanni Tummarello. Searching web data: An entity retrieval and high-performance indexing model. *Web Semantics: Science, Services and Agents on the World Wide Web*, 10(0):33 – 58, 2012. Web-Scale Semantic Information Processing
- Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker. Searching and browsing linked data with SWSE: The semantic web search engine. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):365 – 401, 2011. {JWS} special issue on Semantic Search
- Thanh Tran, Haofen Wang, and Peter Haase. Hermes: Data web search on a pay-as-you-go integration infrastructure. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):189 – 203, 2009. The Web of Data

- T. Härder, B. Mitschang, and H. Schöning. Query processing for complex objects. *Data & Knowledge Engineering*, 7(3):181 – 200, 1992
- Timos K. Sellis, Nick Roussopoulos, and Raymond T. Ng. Efficient compilation of large rule bases using logical access paths. *Information Systems*, 15(1):73 – 84, 1990. Knowledge Engineering
- Benny Kimelfeld and Yehoshua Sagiv. Efficiently enumerating results of keyword search over data graphs. *Information Systems*, 33(45):335 – 359, 2008. Selected Papers from the Tenth International Symposium on Database Programming Languages (DBPL 2005
- Raoul-Gabriel Urma and Alan Mycroft. Source-code queries with graph databases with application to programming language usage and evolution. *Science of Computer Programming*, 97, Part 1(0):127 – 134, 2015. Special Issue on New Ideas and Emerging Results in Understanding Software
- Klara Weiand, Fabian KneiSSl, Wojciech Lobacz, Tim Furche, and François Bry. pest: Fast approximate keyword search in semantic data using eigenvector-based term propagation. *Information Systems*, 37(4):372 – 390, 2012. Semantic Web Data Management
- Xiang Lian, Eugenio De Hoyos, Artem Chebotko, Bin Fu, and Christine Reilly. k-nearest keyword search in RDF graphs. *Web Semantics: Science, Services and Agents on the World Wide Web*, pages 40 – 56, 2013
- Michel Mainguenaud. Constraint-based queries in a geographical database for network facilities. *Computers, Environment and Urban Systems*, 20(2):139 – 151, 1996
- Stewart Greenhill and Svetha Venkatesh. Semantic data modelling and visualisation using noetica. *Data & Knowledge Engineering*, 33(3):241 – 276, 2000
- H.M. Jamil. Design of declarative graph query languages: On the choice between value, pattern and object based representations for graphs. In *IEEE 28th International Conference on Data Engineering Workshops (ICDEW)*, pages 178–185, April 2012
- Hong-Han Shuai, De-Nian Yang, P.S. Yu, Chih-Ya Shen, and Ming-Syan Chen. On pattern preserving graph generation. In *IEEE 13th International Conference on Data Mining (ICDM)*, pages 677–686, Dec 2013
- A. Petermann, M. Junghanns, R. Muller, and E. Rahm. BIIIG: Enabling business intelligence with integrated instance graphs. In *IEEE 30th International Conference on Data Engineering Workshops (ICDEW)*, pages 4–11, March 2014
- A. Morari, V.G. Castellana, D. Haglin, J. Feo, J. Weaver, A. Tumeo, and O. Villa. Accelerating semantic graph databases on commodity clusters. In *IEEE International Conference on Big Data*, pages 768–772, Oct 2013
- Yijun Bei, Zhen Lin, Chen Zhao, and Xiaojun Zhu. HBase system-based distributed framework for searching large graph databases. In *14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 151–156, July 2013
- A. Naseer, L. Laera, and T. Matsutsuka. Enterprise biggraph. In *46th Hawaii International Conference on System Sciences (HICSS)*, pages 1005–1014, Jan 2013
- M. Uddin, R. Stadler, M. Miyazawa, and M. Hayashi. Graph search for cloud network management. In *Network Operations and Management Symposium (NOMS), 2014 IEEE*, pages 1–5, May 2014

- V.G. Castellana, A. Tumeo, O. Villa, D. Haglin, and J. Feo. Composing data parallel code for a SPARQL graph engine. In *International Conference on Social Computing (SocialCom)*, pages 691–699, Sept 2013
- W.E. Moustafa, A. Kimmig, A. Deshpande, and L. Getoor. Subgraph pattern matching over uncertain graphs with identity linkage uncertainty. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 904–915, March 2014
- M. Brocheler, A. Pugliese, and V.S. Subrahmanian. COSI: Cloud oriented subgraph identification in massive social networks. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 248–255, Aug 2010
- A. Cuzzocrea and P. Serafino. A reachability-based theoretical framework for modeling and querying complex probabilistic graph data. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1177–1184, Oct 2012
- Rui Wang and K. Chiu. A stream partitioning approach to processing large scale distributed graph datasets. In *IEEE International Conference on Big Data*, pages 537–542, Oct 2013
- S.R. Spillane, J. Birnbaum, D. Bokser, D. Kemp, A. Labouseur, P.W. Olsen, J. Vijayan, Jeong-Hyon Hwang, and Jun-Weon Yoon. A demonstration of the G* graph database system. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 1356–1359, April 2013
- Yunkai Liu and T.M. Vitolo. Graph data warehouse: Steps to integrating graph databases into the traditional conceptual structure of a data warehouse. In *IEEE International Congress on Big Data (BigData Congress)*, pages 433–434, June 2013
- L. Fegaras. Supporting Bulk Synchronous Parallelism in Map-Reduce queries. In *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*, pages 1068–1077, Nov 2012
- O. Cure, F. Kerdjoudj, Chan Le Duc, M. Lamolle, and D. Faye. On the potential integration of an ontology-based data access approach in NoSQL stores. In *Third International Conference on Emerging Intelligent Data and Web Technologies (EIDWT)*, pages 166–173, Sept 2012
- M. Leida and A. Chu. Distributed SPARQL query answering over RDF data streams. In *IEEE International Congress on Big Data (BigData Congress)*, pages 369–378, June 2013
- Ying Tu and Han-Wei Shen. GraphCharter: Combining browsing with query to explore large semantic graphs. In *Visualization Symposium (PacificVis), 2013 IEEE Pacific*, pages 49–56, Feb 2013
- V. Dritsou, P. Constantopoulos, A. Deligiannakis, and Y. Kotidis. Shortcut selection in RDF databases. In *IEEE 27th International Conference on Data Engineering Workshops (ICDEW)*, pages 194–199, April 2011
- Letao Qi, H.T. Lin, and V. Honavar. Clustering remote RDF data using SPARQL update queries. In *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 236–242, April 2013
- G. Kasneci, M. Ramanath, M. Sozio, F.M. Suchanek, and G. Weikum. STAR: Steiner-tree approximation in relationship graphs. In *IEEE 25th International Conference on Data Engineering - ICDE'09*, pages 868–879, March 2009
- Ruoming Jin, Yang Xiang, Ning Ruan, and David Fuhry. 3-HOP: A high-compression indexing scheme for reachability query. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, pages 813–826, New York, NY, USA, 2009. ACM

- Kai Zeng, Jiacheng Yang, Haixun Wang, Bin Shao, and Zhongyuan Wang. A distributed graph engine for web scale RDF data. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pages 265–276. VLDB Endowment, 2013
- Andrey Kan, Jeffrey Chan, James Bailey, and Christopher Leckie. A query based approach for mining evolving graphs. In *Proceedings of the Eighth Australasian Data Mining Conference - Volume 101*, AusDM '09, pages 139–150, Darlinghurst, Australia, Australia, 2009. Australian Computer Society, Inc
- Hao He, Haixun Wang, Jun Yang, and Philip S. Yu. BLINKS: Ranked keyword searches on graphs. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 305–316, New York, NY, USA, 2007. ACM
- Dritan Bleco and Yannis Kotidis. Business intelligence on complex graph data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, pages 13–20, New York, NY, USA, 2012. ACM
- Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 125–134, New York, NY, USA, 2012. ACM
- Roberto De Virgilio, Antonio Maccioni, and Riccardo Torlone. Converting relational to graph databases. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 1:1–1:6, New York, NY, USA, 2013. ACM
- Dan Suciu. Distributed query evaluation on semistructured data. *ACM Trans. Database Syst.*, 27(1):1–62, March 2002
- Jayanta Mondal and Amol Deshpande. EAGr: Supporting continuous ego-centric aggregate queries over large dynamic graphs. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1335–1346, New York, NY, USA, 2014. ACM
- Zhao Sun, Hongzhi Wang, Haixun Wang, Bin Shao, and Jianzhong Li. Efficient subgraph matching on billion node graphs. *Proc. VLDB Endow.*, 5(9):788–799, May 2012
- Wenlei Xie, Guozhang Wang, David Bindel, Alan Demers, and Johannes Gehrke. Fast iterative graph computation with block updates. *Proc. VLDB Endow.*, 6(14):2014–2025, September 2013
- Ramakrishna Varadarajan, Vagelis Hristidis, Louiqa Raschid, Maria-Esther Vidal, Luis Ibáñez, and Héctor Rodríguez-Drumond. Flexible and efficient querying and ranking on hyperlinked data sources. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 553–564, New York, NY, USA, 2009. ACM
- Federica Mandreoli, Riccardo Martoglia, Giorgio Villani, and Wilma Penzo. Flexible query answering on graph-modeled data. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 216–227, New York, NY, USA, 2009. ACM
- Semih Salihoglu and Jennifer Widom. GPS: A graph processing system. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, SSDBM, pages 22:1–22:12, New York, NY, USA, 2013. ACM
- Hilmi Yildirim, Vineet Chaoji, and Mohammed J. Zaki. GRAIL: A scalable index for reachability queries in very large graphs. *The VLDB Journal*, 21(4):509–534, August 2012

- María Constanza Pabón, Claudia Roncancio, and Martha Millán. Graph data transformations and querying. In *Proceedings of the 2014 International C* Conference on Computer Science & Software Engineering*, C3S2E '14, pages 20:1–20:6, New York, NY, USA, 2014. ACM
- Lei Zou, M. Tamer Özsu, Lei Chen, Xuchuan Shen, Ruizhe Huang, and Dongyan Zhao. gStore: A graph-based SPARQL query engine. *The VLDB Journal*, 23(4):565–590, August 2014
- Wenfei Fan, Xin Wang, and Yinghui Wu. Incremental graph pattern matching. *ACM Trans. Database Syst.*, 38(3):18:1–18:47, September 2013
- Yosi Mass and Yehoshua Sagiv. Language models for keyword search over data graphs. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 363–372, New York, NY, USA, 2012. ACM
- Arijit Khan, Nan Li, Xifeng Yan, Ziyu Guan, Supriyo Chakraborty, and Shu Tao. Neighborhood based fast graph search in large networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 901–912, New York, NY, USA, 2011. ACM
- Saumen Dey, Víctor Cuevas-Vicenttín, Sven Köhler, Eric Gribkoff, Michael Wang, and Bertram Ludäscher. On implementing provenance-aware regular path queries with relational query engines. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 214–223, New York, NY, USA, 2013. ACM
- Yuanyuan Tian, Jignesh M. Patel, Viji Nair, Sebastian Martini, and Matthias Kretzler. Periscope/GQ: A graph querying toolkit. *Proc. VLDB Endow.*, 1(2):1404–1407, August 2008
- Shengqi Yang, Yanan Xie, Yinghui Wu, Tianyu Wu, Huan Sun, Jian Wu, and Xifeng Yan. SLQ: A user-friendly graph querying system. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 893–896, New York, NY, USA, 2014. ACM
- Ciro Cattuto, Marco Quaggiotto, André Panisson, and Alex Averbuch. Time-varying social networks in a graph database: A Neo4J use case. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 11:1–11:6, New York, NY, USA, 2013. ACM
- Padmashree Ravindra, HyeongSik Kim, and Kemafor Anyanwu. To nest or not to nest, when and how much: Representing intermediate results of graph pattern queries in mapreduce based processing. In *Proceedings of the 4th International Workshop on Semantic Web Information Management*, SWIM '12, pages 5:1–5:8, New York, NY, USA, 2012. ACM
- Alan Chappell, Sutanay Choudhury, John Feo, David Haglin, Alessandro Morari, Sumit Purohit, Karen Schuchardt, Antonino Tumeo, Jesse Weaver, and Oreste Villa. Toward a data scalable solution for facilitating discovery of scientific data resources. In *Proceedings of the 2013 International Workshop on Data-Intensive Scalable Computing Systems*, DISCS-2013, pages 55–60, New York, NY, USA, 2013. ACM
- Shengqi Yang, Xifeng Yan, Bo Zong, and Arijit Khan. Towards effective partition management for large graphs. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 517–528, New York, NY, USA, 2012. ACM
- Norbert Martinez-Bazan and David Dominguez-Sal. Using semijoin programs to solve traversal queries in graph databases. In *Proceedings of Workshop on GRAph Data Management Experiences and Systems*, GRADES'14, pages 6:1–6:6, New York, NY, USA, 2014. ACM

.

Referências dos artigos subcategoria Comparação

- Jesse Weaver, Vito Giovanni Castellana, Alessandro Morari, Antonino Tumeo, Sumit Purohit, Alan Chappell, David Haglin, Oreste Villa, Sutanay Choudhury, Karen Schuchardt, and John Feo. Toward a data scalable solution for facilitating discovery of science resources. *Parallel Computing*, 40(10):682 – 696, 2014
- Renaud Delbru, Stephane Campinas, and Giovanni Tummarello. Searching web data: An entity retrieval and high-performance indexing model. *Web Semantics: Science, Services and Agents on the World Wide Web*, 10(0):33 – 58, 2012. Web-Scale Semantic Information Processing
- Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of linked data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14(0):14 – 44, 2012. Special Issue on Dealing with the Messiness of the Web of Data
- A. González-Beltrán, P. Milligan, and P. Sage. Range queries over skip tree graphs. *Computer Communications*, 31(2):358 – 374, 2008. Special Issue: Foundation of Peer-to-Peer Computing
-
- Paul Groth, Antonis Loizou, Alasdair J.G. Gray, Carole Goble, Lee Harland, and Steve Pettifer. Api-centric linked data integration: The open PHACTS discovery platform case study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2014
- Ying Ding. Community detection: Topological vs. topical. *Journal of Informetrics*, 5(4):498 – 514, 2011
- A. Khan, Yinghui Wu, and X. Yan. Emerging graph queries in linked data. In *IEEE 28th International Conference on Data Engineering (ICDE)*, pages 1218–1221, April 2012
- S. Jouili and V. Vansteenberghe. An empirical comparison of graph databases. In *International Conference on Social Computing (SocialCom)*, pages 708–715, Sept 2013
- K. Kaur and R. Rani. Modeling and querying data in NoSQL databases. In *IEEE International Conference on Big Data*, pages 1–7, Oct 2013
- R. Angles. A comparison of current graph database models. In *IEEE 28th International Conference on Data Engineering Workshops (ICDEW)*, pages 171–177, April 2012
- Yi Chen, Wei Wang, and Ziyang Liu. Keyword-based search and exploration on databases. In *IEEE 27th International Conference on Data Engineering (ICDE)*, pages 1380–1383, April 2011
- Renzo Angles, Arnau Prat-Pérez, David Dominguez-Sal, and Josep-Lluis Larriba-Pey. Benchmarking database systems for social network applications. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 15:1–15:7, New York, NY, USA, 2013. ACM
- Florian Holzschuher and René Peinl. Performance of graph query languages: Comparison of cypher, gremlin and native access in Neo4J. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 195–204, New York, NY, USA, 2013. ACM
- Peter T. Wood. Query languages for graph databases. *SIGMOD Rec.*, 41(1):50–60, April 2012
- Pablo Barceló Baeza. Querying graph databases. In *Proceedings of the 32nd Symposium on Principles of Database Systems*, PODS '13, pages 175–188, New York, NY, USA, 2013. ACM

- Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Comput. Surv.*, 40(1):1:1–1:39, February 2008

.

Referências dos artigos subcategoria Experimentos

- Linhong Zhu, Wee Keong Ng, and James Cheng. Structure and attribute index for approximate graph matching in large graphs. *Information Systems*, 36(6):958 – 972, 2011

- Chang-Sup Park and Sungchae Lim. Efficient processing of keyword queries over graph databases for finding effective answers. *Information Processing & Management*, 51(1):42 – 57, 2015

- Thanh Tran, Haofen Wang, and Peter Haase. Hermes: Data web search on a pay-as-you-go integration infrastructure. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):189 – 203, 2009. The Web of Data

- A. González-Beltrán, P. Milligan, and P. Sage. Range queries over skip tree graphs. *Computer Communications*, 31(2):358 – 374, 2008. Special Issue: Foundation of Peer-to-Peer Computing

- Raoul-Gabriel Urma and Alan Mycroft. Source-code queries with graph databases with application to programming language usage and evolution. *Science of Computer Programming*, 97, Part 1(0):127 – 134, 2015. Special Issue on New Ideas and Emerging Results in Understanding Software

- Klara Weiand, Fabian KneiSSl, Wojciech Lobacz, Tim Furche, and François Bry. pest: Fast approximate keyword search in semantic data using eigenvector-based term propagation. *Information Systems*, 37(4):372 – 390, 2012. Semantic Web Data Management

- Xiang Lian, Eugenio De Hoyos, Artem Chebotko, Bin Fu, and Christine Reilly. k-nearest keyword search in RDF graphs. *Web Semantics: Science, Services and Agents on the World Wide Web*, pages 40 – 56, 2013

- Hong-Han Shuai, De-Nian Yang, P.S. Yu, Chih-Ya Shen, and Ming-Syan Chen. On pattern preserving graph generation. In *IEEE 13th International Conference on Data Mining (ICDM)*, pages 677–686, Dec 2013

- A. Petermann, M. Junghanns, R. Muller, and E. Rahm. BIIIG: Enabling business intelligence with integrated instance graphs. In *IEEE 30th International Conference on Data Engineering Workshops (ICDEW)*, pages 4–11, March 2014

- A. Morari, V.G. Castellana, D. Haglin, J. Feo, J. Weaver, A. Tumeo, and O. Villa. Accelerating semantic graph databases on commodity clusters. In *IEEE International Conference on Big Data*, pages 768–772, Oct 2013

- S. Jouili and V. Vansteenberghe. An empirical comparison of graph databases. In *International Conference on Social Computing (SocialCom)*, pages 708–715, Sept 2013

- Yijun Bei, Zhen Lin, Chen Zhao, and Xiaojun Zhu. HBase system-based distributed framework for searching large graph databases. In *14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 151–156, July 2013

- J. Heer and A. Perer. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 51–60, Oct 2011

- W.E. Moustafa, A. Kimmig, A. Deshpande, and L. Getoor. Subgraph pattern matching over uncertain graphs with identity linkage uncertainty. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 904–915, March 2014
- Jiwon Seo, S. Guo, and M.S. Lam. SociaLite: Datalog extensions for efficient social network analysis. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 278–289, April 2013
- Rui Wang and K. Chiu. A stream partitioning approach to processing large scale distributed graph datasets. In *IEEE International Conference on Big Data*, pages 537–542, Oct 2013
- Jun Gao, Jiashuai Zhou, Chang Zhou, and J.X. Yu. GLog: A high level graph analysis system using mapreduce. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 544–555, March 2014
- Ying Tu and Han-Wei Shen. GraphCharter: Combining browsing with query to explore large semantic graphs. In *Visualization Symposium (PacificVis), 2013 IEEE Pacific*, pages 49–56, Feb 2013
- P. Yadav and V. Samala. Benchmarking over a semantic repository. In *Second International Conference on Advanced Computing (ICoAC)*, pages 51–59, Dec 2010
- Kai Zeng, Jiacheng Yang, Haixun Wang, Bin Shao, and Zhongyuan Wang. A distributed graph engine for web scale RDF data. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pages 265–276. VLDB Endowment, 2013
- Andrey Kan, Jeffrey Chan, James Bailey, and Christopher Leckie. A query based approach for mining evolving graphs. In *Proceedings of the Eighth Australasian Data Mining Conference - Volume 101*, AusDM '09, pages 139–150, Darlinghurst, Australia, Australia, 2009. Australian Computer Society, Inc
- Roberto De Virgilio, Antonio Maccioni, and Riccardo Torlone. Converting relational to graph databases. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 1:1–1:6, New York, NY, USA, 2013. ACM
- Jayanta Mondal and Amol Deshpande. EAGr: Supporting continuous ego-centric aggregate queries over large dynamic graphs. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1335–1346, New York, NY, USA, 2014. ACM
- Zhao Sun, Hongzhi Wang, Haixun Wang, Bin Shao, and Jianzhong Li. Efficient subgraph matching on billion node graphs. *Proc. VLDB Endow.*, 5(9):788–799, May 2012
- Vitaliy Liptchinsky, Benjamin Satzger, Rostyslav Zabolotnyi, and Schahram Dustdar. Expressive languages for selecting groups from graph-structured data. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 761–770, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee
- Wenlei Xie, Guozhang Wang, David Bindel, Alan Demers, and Johannes Gehrke. Fast iterative graph computation with block updates. *Proc. VLDB Endow.*, 6(14):2014–2025, September 2013
- Federica Mandreoli, Riccardo Martoglia, Giorgio Villani, and Wilma Penzo. Flexible query answering on graph-modeled data. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 216–227, New York, NY, USA, 2009. ACM

- Semih Salihoglu and Jennifer Widom. GPS: A graph processing system. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, SSDBM, pages 22:1–22:12, New York, NY, USA, 2013. ACM
- Lei Zou, M. Tamer Özsu, Lei Chen, Xuchuan Shen, Ruizhe Huang, and Dongyan Zhao. gStore: A graph-based SPARQL query engine. *The VLDB Journal*, 23(4):565–590, August 2014
- Mohamed Sarwat, Sameh Elnikety, Yuxiong He, and Mohamed F. Mokbel. Horton+: A distributed system for processing declarative reachability queries over partitioned graphs. *Proc. VLDB Endow.*, 6(14):1918–1929, September 2013
- Wenfei Fan, Xin Wang, and Yinghui Wu. Incremental graph pattern matching. *ACM Trans. Database Syst.*, 38(3):18:1–18:47, September 2013
- Arijit Khan, Nan Li, Xifeng Yan, Ziyu Guan, Supriyo Chakraborty, and Shu Tao. Neighborhood based fast graph search in large networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 901–912, New York, NY, USA, 2011. ACM
- Saumen Dey, Víctor Cuevas-Vicenttín, Sven Köhler, Eric Gribkoff, Michael Wang, and Bertram Ludäscher. On implementing provenance-aware regular path queries with relational query engines. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 214–223, New York, NY, USA, 2013. ACM
- Juan P. Cedeño and K. Selçuk Candan. R2DF framework for ranked path queries over weighted RDF graphs. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 40:1–40:12, New York, NY, USA, 2011. ACM
- Shengqi Yang, Xifeng Yan, Bo Zong, and Arijit Khan. Towards effective partition management for large graphs. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 517–528, New York, NY, USA, 2012. ACM
.

Referências dos artigos subcategoria Ferramenta

- Stewart Greenhill and Svetha Venkatesh. Semantic data modelling and visualisation using noetica. *Data & Knowledge Engineering*, 33(3):241 – 276, 2000
- R. Ronen and O. Shmueli. SoQL: A language for querying and creating data in social networks. In *ICDE'09. IEEE 25th International Conference on Data Engineering*, pages 1595–1602, March 2009
- A. Naseer, L. Laera, and T. Matsutsuka. Enterprise biggraph. In *46th Hawaii International Conference on System Sciences (HICSS)*, pages 1005–1014, Jan 2013
- Ying Tu and Han-Wei Shen. GraphCharter: Combining browsing with query to explore large semantic graphs. In *Visualization Symposium (PacificVis), 2013 IEEE Pacific*, pages 49–56, Feb 2013
- Hilmi Yildirim, Vineet Chaoji, and Mohammed J. Zaki. GRAIL: A scalable index for reachability queries in very large graphs. *The VLDB Journal*, 21(4):509–534, August 2012
- Yuanyuan Tian, Jignesh M. Patel, Viji Nair, Sebastian Martini, and Matthias Kretzler. Periscope/GQ: A graph querying toolkit. *Proc. VLDB Endow.*, 1(2):1404–1407, August 2008
- Jingren Zhou, Nicolas Bruno, Ming-Chuan Wu, Per-Ake Larson, Ronnie Chaiken, and Darren Shakib. SCOPE: Parallel databases meet MapReduce. *The VLDB Journal*, 21(5):611–636, October 2012

Tabela 5: Categoria: tipo de problema.

Referências dos artigos subcategoria Processamento de consultas
- Sherif Sakr, Sameh Elnikety, and Yuxiong He. Hybrid query execution engine for large attributed graphs. *Information Systems*, 41(0):45 – 73, 2014
- Jesse Weaver, Vito Giovanni Castellana, Alessandro Morari, Antonino Tumeo, Sumit Purohit, Alan Chappell, David Haglin, Oreste Villa, Sutanay Choudhury, Karen Schuchardt, and John Feo. Toward a data scalable solution for facilitating discovery of science resources. *Parallel Computing*, 40(10):682 – 696, 2014
- Linhong Zhu, Wee Keong Ng, and James Cheng. Structure and attribute index for approximate graph matching in large graphs. *Information Systems*, 36(6):958 – 972, 2011
- Chang-Sup Park and Sungchae Lim. Efficient processing of keyword queries over graph databases for finding effective answers. *Information Processing & Management*, 51(1):42 – 57, 2015
- Renaud Delbru, Stephane Campinas, and Giovanni Tummarello. Searching web data: An entity retrieval and high-performance indexing model. *Web Semantics: Science, Services and Agents on the World Wide Web*, 10(0):33 – 58, 2012. Web-Scale Semantic Information Processing
- Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker. Searching and browsing linked data with SWSE: The semantic web search engine. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):365 – 401, 2011. {JWS} special issue on Semantic Search
- Thanh Tran, Haofen Wang, and Peter Haase. Hermes: Data web search on a pay-as-you-go integration infrastructure. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):189 – 203, 2009. The Web of Data
- T. Härder, B. Mitschang, and H. Schöning. Query processing for complex objects. *Data & Knowledge Engineering*, 7(3):181 – 200, 1992
- Timos K. Sellis, Nick Roussopoulos, and Raymond T. Ng. Efficient compilation of large rule bases using logical access paths. *Information Systems*, 15(1):73 – 84, 1990. Knowledge Engineering
- Raoul-Gabriel Urma and Alan Mycroft. Source-code queries with graph databases with application to programming language usage and evolution. *Science of Computer Programming*, 97, Part 1(0):127 – 134, 2015. Special Issue on New Ideas and Emerging Results in Understanding Software
- Klara Weiand, Fabian KneiSSl, Wojciech Lobacz, Tim Furche, and François Bry. pest: Fast approximate keyword search in semantic data using eigenvector-based term propagation. *Information Systems*, 37(4):372 – 390, 2012. Semantic Web Data Management
- Xiang Lian, Eugenio De Hoyos, Artem Chebotko, Bin Fu, and Christine Reilly. k-nearest keyword search in RDF graphs. *Web Semantics: Science, Services and Agents on the World Wide Web*, pages 40 – 56, 2013
- Michel Mainguenaud. Constraint-based queries in a geographical database for network facilities. *Computers, Environment and Urban Systems*, 20(2):139 – 151, 1996
- Stewart Greenhill and Svetha Venkatesh. Semantic data modelling and visualisation using noetica. *Data & Knowledge Engineering*, 33(3):241 – 276, 2000

- G. Karvounarakis, A. Magganaraki, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, and K. Tolle. Querying the semantic web with RQL. *Computer Networks*, 42(5):617 – 640, 2003. The Semantic Web: an evolution for a revolution
- H.M. Jamil. Design of declarative graph query languages: On the choice between value, pattern and object based representations for graphs. In *IEEE 28th International Conference on Data Engineering Workshops (ICDEW)*, pages 178–185, April 2012
- A. Khan, Yinghui Wu, and X. Yan. Emerging graph queries in linked data. In *IEEE 28th International Conference on Data Engineering (ICDE)*, pages 1218–1221, April 2012
- A. Petermann, M. Junghanns, R. Muller, and E. Rahm. BIIIG: Enabling business intelligence with integrated instance graphs. In *IEEE 30th International Conference on Data Engineering Workshops (ICDEW)*, pages 4–11, March 2014
- R. Ronen and O. Shmueli. SoQL: A language for querying and creating data in social networks. In *ICDE'09. IEEE 25th International Conference on Data Engineering*, pages 1595–1602, March 2009
- Yijun Bei, Zhen Lin, Chen Zhao, and Xiaojun Zhu. HBase system-based distributed framework for searching large graph databases. In *14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 151–156, July 2013
- J. Heer and A. Perer. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 51–60, Oct 2011
- A. Naseer, L. Laera, and T. Matsutsuka. Enterprise biggraph. In *46th Hawaii International Conference on System Sciences (HICSS)*, pages 1005–1014, Jan 2013
- M. Uddin, R. Stadler, M. Miyazawa, and M. Hayashi. Graph search for cloud network management. In *Network Operations and Management Symposium (NOMS), 2014 IEEE*, pages 1–5, May 2014
- V.G. Castellana, A. Tumeo, O. Villa, D. Haglin, and J. Feo. Composing data parallel code for a SPARQL graph engine. In *International Conference on Social Computing (SocialCom)*, pages 691–699, Sept 2013
- W.E. Moustafa, A. Kimmig, A. Deshpande, and L. Getoor. Subgraph pattern matching over uncertain graphs with identity linkage uncertainty. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 904–915, March 2014
- Jiwon Seo, S. Guo, and M.S. Lam. SociaLite: Datalog extensions for efficient social network analysis. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 278–289, April 2013
- M. Brocheler, A. Pugliese, and V.S. Subrahmanian. COSI: Cloud oriented subgraph identification in massive social networks. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 248–255, Aug 2010
- A. Cuzzocrea and P. Serafino. A reachability-based theoretical framework for modeling and querying complex probabilistic graph data. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1177–1184, Oct 2012
- Rui Wang and K. Chiu. A stream partitioning approach to processing large scale distributed graph datasets. In *IEEE International Conference on Big Data*, pages 537–542, Oct 2013

- K. Kaur and R. Rani. Modeling and querying data in NoSQL databases. In *IEEE International Conference on Big Data*, pages 1–7, Oct 2013
- S.R. Spillane, J. Birnbaum, D. Bokser, D. Kemp, A. Labouseur, P.W. Olsen, J. Vijayan, Jeong-Hyon Hwang, and Jun-Weon Yoon. A demonstration of the G* graph database system. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 1356–1359, April 2013
- R. Angles. A comparison of current graph database models. In *IEEE 28th International Conference on Data Engineering Workshops (ICDEW)*, pages 171–177, April 2012
- Yunkai Liu and T.M. Vitolo. Graph data warehouse: Steps to integrating graph databases into the traditional conceptual structure of a data warehouse. In *IEEE International Congress on Big Data (BigData Congress)*, pages 433–434, June 2013
- C. Raguenaud, J. Kennedy, and P.J. Barclay. The Prometheus taxonomic database. In *Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pages 63–70, 2000
- Jun Gao, Jiashuai Zhou, Chang Zhou, and J.X. Yu. GLog: A high level graph analysis system using mapreduce. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 544–555, March 2014
- L. Fegaras. Supporting Bulk Synchronous Parallelism in Map-Reduce queries. In *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*, pages 1068–1077, Nov 2012
- O. Cure, F. Kerdjoudj, Chan Le Duc, M. Lamolle, and D. Faye. On the potential integration of an ontology-based data access approach in NoSQL stores. In *Third International Conference on Emerging Intelligent Data and Web Technologies (EIDWT)*, pages 166–173, Sept 2012
- M. Leida and A. Chu. Distributed SPARQL query answering over RDF data streams. In *IEEE International Congress on Big Data (BigData Congress)*, pages 369–378, June 2013
- Ying Tu and Han-Wei Shen. GraphCharter: Combining browsing with query to explore large semantic graphs. In *Visualization Symposium (PacificVis), 2013 IEEE Pacific*, pages 49–56, Feb 2013
- P. Yadav and V. Samala. Benchmarking over a semantic repository. In *Second International Conference on Advanced Computing (ICoAC)*, pages 51–59, Dec 2010
- Taekyong Lee, Lei Sheng, T. Bozkaya, N.H. Balkir, Z. Meral Ozsoyoglu, and Z.M. Ozsoyoglu. Querying multimedia presentations based on content. *IEEE Transactions on Knowledge and Data Engineering*, 11(3):361–385, May 1999
- S. Dar and R. Agrawal. Extending SQL with generalized transitive closure. *IEEE Transactions on Knowledge and Data Engineering*, 5(5):799–812, Oct 1993
- V. Dritsou, P. Constantopoulos, A. Deligiannakis, and Y. Kotidis. Shortcut selection in RDF databases. In *IEEE 27th International Conference on Data Engineering Workshops (ICDEW)*, pages 194–199, April 2011
- Letao Qi, H.T. Lin, and V. Honavar. Clustering remote RDF data using SPARQL update queries. In *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 236–242, April 2013
- G. Kasneci, M. Ramanath, M. Sozio, F.M. Suchanek, and G. Weikum. STAR: Steiner-tree approximation in relationship graphs. In *IEEE 25th International Conference on Data Engineering - ICDE'09*, pages 868–879, March 2009

- M. Jarrar and M.D. Dikaiakos. Querying the data web: The MashQL approach. *Internet Computing, IEEE*, 14(3):58–67, May 2010
- Ruoming Jin, Yang Xiang, Ning Ruan, and David Fuhry. 3-HOP: A high-compression indexing scheme for reachability query. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, pages 813–826, New York, NY, USA, 2009. ACM
- Kai Zeng, Jiacheng Yang, Haixun Wang, Bin Shao, and Zhongyuan Wang. A distributed graph engine for web scale RDF data. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pages 265–276. VLDB Endowment, 2013
- Andrey Kan, Jeffrey Chan, James Bailey, and Christopher Leckie. A query based approach for mining evolving graphs. In *Proceedings of the Eighth Australasian Data Mining Conference - Volume 101*, AusDM '09, pages 139–150, Darlinghurst, Australia, Australia, 2009. Australian Computer Society, Inc
- Anton Dries, Siegfried Nijssen, and Luc De Raedt. A query language for analyzing networks. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 485–494, New York, NY, USA, 2009. ACM
- Anton Dries and Siegfried Nijssen. Analyzing graph databases by aggregate queries. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, MLG '10, pages 37–45, New York, NY, USA, 2010. ACM
- Fan Wang and Gagan Agrawal. Answering complex structured queries over the deep web. In *Proceedings of the 15th Symposium on International Database Engineering & Applications*, IDEAS '11, pages 115–123, New York, NY, USA, 2011. ACM
- Renzo Angles, Arnau Prat-Pérez, David Dominguez-Sal, and Josep-Lluis Larriba-Pey. Benchmarking database systems for social network applications. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 15:1–15:7, New York, NY, USA, 2013. ACM
- Hao He, Haixun Wang, Jun Yang, and Philip S. Yu. BLINKS: Ranked keyword searches on graphs. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 305–316, New York, NY, USA, 2007. ACM
- Dritan Bleco and Yannis Kotidis. Business intelligence on complex graph data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, pages 13–20, New York, NY, USA, 2012. ACM
- Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 125–134, New York, NY, USA, 2012. ACM
- Dan Suciu. Distributed query evaluation on semistructured data. *ACM Trans. Database Syst.*, 27(1):1–62, March 2002
- Jayanta Mondal and Amol Deshpande. EAGr: Supporting continuous ego-centric aggregate queries over large dynamic graphs. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1335–1346, New York, NY, USA, 2014. ACM
- Zhao Sun, Hongzhi Wang, Haixun Wang, Bin Shao, and Jianzhong Li. Efficient subgraph matching on billion node graphs. *Proc. VLDB Endow.*, 5(9):788–799, May 2012

- M. P. Consens and A. O. Mendelzon. Expressing structural hypertext queries in graphlog. In *Proceedings of the Second Annual ACM Conference on Hypertext*, HYPERTEXT '89, pages 269–292, New York, NY, USA, 1989. ACM
- Vitaliy Liptchinsky, Benjamin Satzger, Rostyslav Zabolotnyi, and Schahram Dustdar. Expressive languages for selecting groups from graph-structured data. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 761–770, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee
- Mirjana Mazuran, Edoardo Serra, and Carlo Zaniolo. Extending the power of datalog recursion. *The VLDB Journal*, 22(4):471–493, August 2013
- Wenlei Xie, Guozhang Wang, David Bindel, Alan Demers, and Johannes Gehrke. Fast iterative graph computation with block updates. *Proc. VLDB Endow.*, 6(14):2014–2025, September 2013
- Ramakrishna Varadarajan, Vagelis Hristidis, Louiqa Raschid, Maria-Esther Vidal, Luis Ibáñez, and Héctor Rodríguez-Drumond. Flexible and efficient querying and ranking on hyperlinked data sources. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 553–564, New York, NY, USA, 2009. ACM
- Federica Mandreoli, Riccardo Martoglia, Giorgio Villani, and Wilma Penzo. Flexible query answering on graph-modeled data. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 216–227, New York, NY, USA, 2009. ACM
- Semih Salihoglu and Jennifer Widom. GPS: A graph processing system. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, SSDBM, pages 22:1–22:12, New York, NY, USA, 2013. ACM
- Hilmi Yildirim, Vineet Chaoji, and Mohammed J. Zaki. GRAIL: A scalable index for reachability queries in very large graphs. *The VLDB Journal*, 21(4):509–534, August 2012
- María Constanza Pabón, Claudia Roncancio, and Martha Millán. Graph data transformations and querying. In *Proceedings of the 2014 International C* Conference on Computer Science & Software Engineering*, C3S2E '14, pages 20:1–20:6, New York, NY, USA, 2014. ACM
- Lei Zou, M. Tamer Özsu, Lei Chen, Xuchuan Shen, Ruizhe Huang, and Dongyan Zhao. gStore: A graph-based SPARQL query engine. *The VLDB Journal*, 23(4):565–590, August 2014
- Mohamed Sarwat, Sameh Elnikety, Yuxiong He, and Mohamed F. Mokbel. Horton+: A distributed system for processing declarative reachability queries over partitioned graphs. *Proc. VLDB Endow.*, 6(14):1918–1929, September 2013
- Wenfei Fan, Xin Wang, and Yinghui Wu. Incremental graph pattern matching. *ACM Trans. Database Syst.*, 38(3):18:1–18:47, September 2013
- Yosi Mass and Yehoshua Sagiv. Language models for keyword search over data graphs. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 363–372, New York, NY, USA, 2012. ACM
- Arijit Khan, Nan Li, Xifeng Yan, Ziyu Guan, Supriyo Chakraborty, and Shu Tao. Neighborhood based fast graph search in large networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 901–912, New York, NY, USA, 2011. ACM

- Saumen Dey, Víctor Cuevas-Vicenttín, Sven Köhler, Eric Gribkoff, Michael Wang, and Bertram Ludäscher. On implementing provenance-aware regular path queries with relational query engines. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 214–223, New York, NY, USA, 2013. ACM
- Amarnath Gupta and Simone Santini. On querying OBO ontologies using a DAG pattern query language. In *Proceedings of the Third International Conference on Data Integration in the Life Sciences*, DILS'06, pages 152–167, Berlin, Heidelberg, 2006. Springer-Verlag
- Florian Holzschuher and René Peinl. Performance of graph query languages: Comparison of cypher, gremlin and native access in Neo4J. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 195–204, New York, NY, USA, 2013. ACM
- Yuanyuan Tian, Jignesh M. Patel, Viji Nair, Sebastian Martini, and Matthias Kretzler. Periscope/GQ: A graph querying toolkit. *Proc. VLDB Endow.*, 1(2):1404–1407, August 2008
- Peter T. Wood. Query languages for graph databases. *SIGMOD Rec.*, 41(1):50–60, April 2012
- Yerach Doytsher, Ben Galon, and Yaron Kanza. Querying geo-social data by bridging spatial networks and social networks. In *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, LBSN '10, pages 39–46, New York, NY, USA, 2010. ACM
- Pablo Barceló Baeza. Querying graph databases. In *Proceedings of the 32nd Symposium on Principles of Database Systems*, PODS '13, pages 175–188, New York, NY, USA, 2013. ACM
- Juan P. Cedeño and K. Selçuk Candan. R2DF framework for ranked path queries over weighted RDF graphs. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 40:1–40:12, New York, NY, USA, 2011. ACM
- Jingren Zhou, Nicolas Bruno, Ming-Chuan Wu, Per-Ake Larson, Ronnie Chaiken, and Darren Shakib. SCOPE: Parallel databases meet MapReduce. *The VLDB Journal*, 21(5):611–636, October 2012
- Shengqi Yang, Yanan Xie, Yinghui Wu, Tianyu Wu, Huan Sun, Jian Wu, and Xifeng Yan. SLQ: A user-friendly graph querying system. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 893–896, New York, NY, USA, 2014. ACM
- Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Comput. Surv.*, 40(1):1:1–1:39, February 2008
- Ciro Cattuto, Marco Quaggiotto, André Panisson, and Alex Averbuch. Time-varying social networks in a graph database: A Neo4J use case. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 11:1–11:6, New York, NY, USA, 2013. ACM
- Padmashree Ravindra, HyeongSik Kim, and Kemafor Anyanwu. To nest or not to nest, when and how much: Representing intermediate results of graph pattern queries in mapreduce based processing. In *Proceedings of the 4th International Workshop on Semantic Web Information Management*, SWIM '12, pages 5:1–5:8, New York, NY, USA, 2012. ACM
- Alan Chappell, Sutanay Choudhury, John Feo, David Haglin, Alessandro Morari, Sumit Purohit, Karen Schuchardt, Antonino Tumeo, Jesse Weaver, and Oreste Villa. Toward a data scalable solution for facilitating discovery of scientific data resources. In *Proceedings of the 2013 International Workshop on Data-Intensive Scalable Computing Systems*, DISCS-2013, pages 55–60, New York, NY, USA, 2013. ACM

- Shengqi Yang, Xifeng Yan, Bo Zong, and Arijit Khan. Towards effective partition management for large graphs. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 517–528, New York, NY, USA, 2012. ACM
- Michael Curtiss, Iain Becker, Tudor Bosman, Sergey Doroshenko, Lucian Grijincu, Tom Jackson, Sandhya Kunnatur, Soren Lassen, Philip Pronin, Sriram Sankar, Guanghao Shen, Gintaras Woss, Chao Yang, and Ning Zhang. Unicorn: A system for searching the social graph. *Proc. VLDB Endow.*, 6(11):1150–1161, August 2013
- Norbert Martinez-Bazan and David Dominguez-Sal. Using semijoin programs to solve traversal queries in graph databases. In *Proceedings of Workshop on GRAph Data Management Experiences and Systems*, GRADES'14, pages 6:1–6:6, New York, NY, USA, 2014. ACM
.

Referências dos artigos subcategoria Grandes grafos

- Sherif Sakr, Sameh Elnikety, and Yuxiong He. Hybrid query execution engine for large attributed graphs. *Information Systems*, 41(0):45 – 73, 2014
- Jesse Weaver, Vito Giovanni Castellana, Alessandro Morari, Antonino Tumeo, Sumit Purohit, Alan Chappell, David Haglin, Oreste Villa, Sutanay Choudhury, Karen Schuchardt, and John Feo. Toward a data scalable solution for facilitating discovery of science resources. *Parallel Computing*, 40(10):682 – 696, 2014
- Linhong Zhu, Wee Keong Ng, and James Cheng. Structure and attribute index for approximate graph matching in large graphs. *Information Systems*, 36(6):958 – 972, 2011
- Chang-Sup Park and Sungchae Lim. Efficient processing of keyword queries over graph databases for finding effective answers. *Information Processing & Management*, 51(1):42 – 57, 2015
- Renaud Delbru, Stephane Campinas, and Giovanni Tummarello. Searching web data: An entity retrieval and high-performance indexing model. *Web Semantics: Science, Services and Agents on the World Wide Web*, 10(0):33 – 58, 2012. Web-Scale Semantic Information Processing
- Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker. Searching and browsing linked data with SWSE: The semantic web search engine. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):365 – 401, 2011. {JWS} special issue on Semantic Search
- Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of linked data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14(0):14 – 44, 2012. Special Issue on Dealing with the Messiness of the Web of Data
- Timos K. Sellis, Nick Roussopoulos, and Raymond T. Ng. Efficient compilation of large rule bases using logical access paths. *Information Systems*, 15(1):73 – 84, 1990. Knowledge Engineering
-
- Klara Weiand, Fabian KneiSSl, Wojciech Lobacz, Tim Furche, and François Bry. pest: Fast approximate keyword search in semantic data using eigenvector-based term propagation. *Information Systems*, 37(4):372 – 390, 2012. Semantic Web Data Management

- Paul Groth, Antonis Loizou, Alasdair J.G. Gray, Carole Goble, Lee Harland, and Steve Pettifer. Api-centric linked data integration: The open PHACTS discovery platform case study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2014
- G. Karvounarakis, A. Magganaraki, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, and K. Tolle. Querying the semantic web with RQL. *Computer Networks*, 42(5):617 – 640, 2003. The Semantic Web: an evolution for a revolution
- Ying Ding. Community detection: Topological vs. topical. *Journal of Informetrics*, 5(4):498 – 514, 2011
- Hong-Han Shuai, De-Nian Yang, P.S. Yu, Chih-Ya Shen, and Ming-Syan Chen. On pattern preserving graph generation. In *IEEE 13th International Conference on Data Mining (ICDM)*, pages 677–686, Dec 2013
- R. Ronen and O. Shmueli. SoQL: A language for querying and creating data in social networks. In *ICDE'09. IEEE 25th International Conference on Data Engineering*, pages 1595–1602, March 2009
- Yijun Bei, Zhen Lin, Chen Zhao, and Xiaojun Zhu. HBase system-based distributed framework for searching large graph databases. In *14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 151–156, July 2013
- J. Heer and A. Perer. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 51–60, Oct 2011
- A. Naseer, L. Laera, and T. Matsutsuka. Enterprise biggraph. In *46th Hawaii International Conference on System Sciences (HICSS)*, pages 1005–1014, Jan 2013
- V.G. Castellana, A. Tumeo, O. Villa, D. Haglin, and J. Feo. Composing data parallel code for a SPARQL graph engine. In *International Conference on Social Computing (SocialCom)*, pages 691–699, Sept 2013
- M. Brocheler, A. Pugliese, and V.S. Subrahmanian. COSI: Cloud oriented subgraph identification in massive social networks. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 248–255, Aug 2010
- Rui Wang and K. Chiu. A stream partitioning approach to processing large scale distributed graph datasets. In *IEEE International Conference on Big Data*, pages 537–542, Oct 2013
- S.R. Spillane, J. Birnbaum, D. Bokser, D. Kemp, A. Labouseur, P.W. Olsen, J. Vijayan, Jeong-Hyon Hwang, and Jun-Weon Yoon. A demonstration of the G* graph database system. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 1356–1359, April 2013
- Jun Gao, Jiashuai Zhou, Chang Zhou, and J.X. Yu. GLog: A high level graph analysis system using mapreduce. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 544–555, March 2014
- Ying Tu and Han-Wei Shen. GraphCharter: Combining browsing with query to explore large semantic graphs. In *Visualization Symposium (PacificVis), 2013 IEEE Pacific*, pages 49–56, Feb 2013
- P. Yadav and V. Samala. Benchmarking over a semantic repository. In *Second International Conference on Advanced Computing (ICoAC)*, pages 51–59, Dec 2010

- V. Dritsou, P. Constantopoulos, A. Deligiannakis, and Y. Kotidis. Shortcut selection in RDF databases. In *IEEE 27th International Conference on Data Engineering Workshops (ICDEW)*, pages 194–199, April 2011
- Letao Qi, H.T. Lin, and V. Honavar. Clustering remote RDF data using SPARQL update queries. In *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 236–242, April 2013
- G. Kasneci, M. Ramanath, M. Sozio, F.M. Suchanek, and G. Weikum. STAR: Steiner-tree approximation in relationship graphs. In *IEEE 25th International Conference on Data Engineering - ICDE'09*, pages 868–879, March 2009
- M. Jarrar and M.D. Dikaiakos. Querying the data web: The MashQL approach. *Internet Computing, IEEE*, 14(3):58–67, May 2010
- Ruoming Jin, Yang Xiang, Ning Ruan, and David Fuhry. 3-HOP: A high-compression indexing scheme for reachability query. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, pages 813–826, New York, NY, USA, 2009. ACM
- Andrey Kan, Jeffrey Chan, James Bailey, and Christopher Leckie. A query based approach for mining evolving graphs. In *Proceedings of the Eighth Australasian Data Mining Conference - Volume 101*, AusDM '09, pages 139–150, Darlinghurst, Australia, Australia, 2009. Australian Computer Society, Inc
- Anton Dries, Siegfried Nijssen, and Luc De Raedt. A query language for analyzing networks. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 485–494, New York, NY, USA, 2009. ACM
- Dritan Bleco and Yannis Kotidis. Business intelligence on complex graph data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, pages 13–20, New York, NY, USA, 2012. ACM
- Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 125–134, New York, NY, USA, 2012. ACM
- Jayanta Mondal and Amol Deshpande. EAGr: Supporting continuous ego-centric aggregate queries over large dynamic graphs. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1335–1346, New York, NY, USA, 2014. ACM
- Zhao Sun, Hongzhi Wang, Haixun Wang, Bin Shao, and Jianzhong Li. Efficient subgraph matching on billion node graphs. *Proc. VLDB Endow.*, 5(9):788–799, May 2012
- Wenlei Xie, Guozhang Wang, David Bindel, Alan Demers, and Johannes Gehrke. Fast iterative graph computation with block updates. *Proc. VLDB Endow.*, 6(14):2014–2025, September 2013
- Semih Salihoglu and Jennifer Widom. GPS: A graph processing system. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, SSDBM, pages 22:1–22:12, New York, NY, USA, 2013. ACM
- Hilmi Yildirim, Vineet Chaoji, and Mohammed J. Zaki. GRAIL: A scalable index for reachability queries in very large graphs. *The VLDB Journal*, 21(4):509–534, August 2012
- Lei Zou, M. Tamer Özsu, Lei Chen, Xuchuan Shen, Ruizhe Huang, and Dongyan Zhao. gStore: A graph-based SPARQL query engine. *The VLDB Journal*, 23(4):565–590, August 2014

- Mohamed Sarwat, Sameh Elnikety, Yuxiong He, and Mohamed F. Mokbel. Horton+: A distributed system for processing declarative reachability queries over partitioned graphs. *Proc. VLDB Endow.*, 6(14):1918–1929, September 2013
- Wenfei Fan, Xin Wang, and Yinghui Wu. Incremental graph pattern matching. *ACM Trans. Database Syst.*, 38(3):18:1–18:47, September 2013
- Yuanyuan Tian, Jignesh M. Patel, Viji Nair, Sebastian Martini, and Matthias Kretzler. Periscope/GQ: A graph querying toolkit. *Proc. VLDB Endow.*, 1(2):1404–1407, August 2008
- Shengqi Yang, Yanan Xie, Yinghui Wu, Tianyu Wu, Huan Sun, Jian Wu, and Xifeng Yan. SLQ: A user-friendly graph querying system. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 893–896, New York, NY, USA, 2014. ACM
- Alan Chappell, Sutanay Choudhury, John Feo, David Haglin, Alessandro Morari, Sumit Purohit, Karen Schuchardt, Antonino Tumeo, Jesse Weaver, and Oreste Villa. Toward a data scalable solution for facilitating discovery of scientific data resources. In *Proceedings of the 2013 International Workshop on Data-Intensive Scalable Computing Systems*, DISCS-2013, pages 55–60, New York, NY, USA, 2013. ACM
- Shengqi Yang, Xifeng Yan, Bo Zong, and Arijit Khan. Towards effective partition management for large graphs. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 517–528, New York, NY, USA, 2012. ACM
- Michael Curtiss, Iain Becker, Tudor Bosman, Sergey Doroshenko, Lucian Grijincu, Tom Jackson, Sandhya Kunnatur, Soren Lassen, Philip Pronin, Sriram Sankar, Guanghao Shen, Gintaras Woss, Chao Yang, and Ning Zhang. Unicorn: A system for searching the social graph. *Proc. VLDB Endow.*, 6(11):1150–1161, August 2013

.

Referências dos artigos subcategoria Funções de agregação

- A. González-Beltrán, P. Milligan, and P. Sage. Range queries over skip tree graphs. *Computer Communications*, 31(2):358 – 374, 2008. Special Issue: Foundation of Peer-to-Peer Computing
- Michel Mainguenaud. Constraint-based queries in a geographical database for network facilities. *Computers, Environment and Urban Systems*, 20(2):139 – 151, 1996
- J. Heer and A. Perer. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 51–60, Oct 2011
- Jiwon Seo, S. Guo, and M.S. Lam. SociaLite: Datalog extensions for efficient social network analysis. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 278–289, April 2013
- S. Dar and R. Agrawal. Extending SQL with generalized transitive closure. *IEEE Transactions on Knowledge and Data Engineering*, 5(5):799–812, Oct 1993
- Anton Dries and Siegfried Nijssen. Analyzing graph databases by aggregate queries. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, MLG '10, pages 37–45, New York, NY, USA, 2010. ACM
- Fan Wang and Gagan Agrawal. Answering complex structured queries over the deep web. In *Proceedings of the 15th Symposium on International Database Engineering & Applications*, IDEAS '11, pages 115–123, New York, NY, USA, 2011. ACM

- Dritan Bleco and Yannis Kotidis. Business intelligence on complex graph data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, pages 13–20, New York, NY, USA, 2012. ACM
- Jayanta Mondal and Amol Deshpande. EAGr: Supporting continuous ego-centric aggregate queries over large dynamic graphs. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1335–1346, New York, NY, USA, 2014. ACM
- M. P. Consens and A. O. Mendelzon. Expressing structural hypertext queries in graphlog. In *Proceedings of the Second Annual ACM Conference on Hypertext*, HYPERTEXT '89, pages 269–292, New York, NY, USA, 1989. ACM
- Lei Zou, M. Tamer Özsu, Lei Chen, Xuchuan Shen, Ruizhe Huang, and Dongyan Zhao. gStore: A graph-based SPARQL query engine. *The VLDB Journal*, 23(4):565–590, August 2014
.

Referências dos artigos subcategoria Palavras-chave

- Chang-Sup Park and Sungchae Lim. Efficient processing of keyword queries over graph databases for finding effective answers. *Information Processing & Management*, 51(1):42 – 57, 2015
- Renaud Delbru, Stephane Campinas, and Giovanni Tummarello. Searching web data: An entity retrieval and high-performance indexing model. *Web Semantics: Science, Services and Agents on the World Wide Web*, 10(0):33 – 58, 2012. Web-Scale Semantic Information Processing
- Thanh Tran, Haofen Wang, and Peter Haase. Hermes: Data web search on a pay-as-you-go integration infrastructure. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):189 – 203, 2009. The Web of Data
- Benny Kimelfeld and Yehoshua Sagiv. Efficiently enumerating results of keyword search over data graphs. *Information Systems*, 33(45):335 – 359, 2008. Selected Papers from the Tenth International Symposium on Database Programming Languages (DBPL 2005
- Xiang Lian, Eugenio De Hoyos, Artem Chebotko, Bin Fu, and Christine Reilly. k-nearest keyword search in RDF graphs. *Web Semantics: Science, Services and Agents on the World Wide Web*, pages 40 – 56, 2013
- M. Uddin, R. Stadler, M. Miyazawa, and M. Hayashi. Graph search for cloud network management. In *Network Operations and Management Symposium (NOMS), 2014 IEEE*, pages 1–5, May 2014
- Yi Chen, Wei Wang, and Ziyang Liu. Keyword-based search and exploration on databases. In *IEEE 27th International Conference on Data Engineering (ICDE)*, pages 1380–1383, April 2011
- Hao He, Haixun Wang, Jun Yang, and Philip S. Yu. BLINKS: Ranked keyword searches on graphs. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 305–316, New York, NY, USA, 2007. ACM
- Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 125–134, New York, NY, USA, 2012. ACM

- Yosi Mass and Yehoshua Sagiv. Language models for keyword search over data graphs. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 363–372, New York, NY, USA, 2012. ACM
- Shengqi Yang, Yanan Xie, Yinghui Wu, Tianyu Wu, Huan Sun, Jian Wu, and Xifeng Yan. SLQ: A user-friendly graph querying system. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 893–896, New York, NY, USA, 2014. ACM

.

Referências dos artigos subcategoria Top-k

- Chang-Sup Park and Sungchae Lim. Efficient processing of keyword queries over graph databases for finding effective answers. *Information Processing & Management*, 51(1):42 – 57, 2015
- A. Cuzzocrea and P. Serafino. A reachability-based theoretical framework for modeling and querying complex probabilistic graph data. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1177–1184, Oct 2012
- V. Dritsou, P. Constantopoulos, A. Deligiannakis, and Y. Kotidis. Shortcut selection in RDF databases. In *IEEE 27th International Conference on Data Engineering Workshops (ICDEW)*, pages 194–199, April 2011
- G. Kasneci, M. Ramanath, M. Sozio, F.M. Suchanek, and G. Weikum. STAR: Steiner-tree approximation in relationship graphs. In *IEEE 25th International Conference on Data Engineering - ICDE'09*, pages 868–879, March 2009
- Anton Dries and Siegfried Nijssen. Analyzing graph databases by aggregate queries. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, MLG '10, pages 37–45, New York, NY, USA, 2010. ACM
- Hao He, Haixun Wang, Jun Yang, and Philip S. Yu. BLINKS: Ranked keyword searches on graphs. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 305–316, New York, NY, USA, 2007. ACM
- Ramakrishna Varadarajan, Vagelis Hristidis, Louiqa Raschid, Maria-Esther Vidal, Luis Ibáñez, and Héctor Rodríguez-Drumond. Flexible and efficient querying and ranking on hyperlinked data sources. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 553–564, New York, NY, USA, 2009. ACM
- Federica Mandreoli, Riccardo Martoglia, Giorgio Villani, and Wilma Penzo. Flexible query answering on graph-modeled data. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 216–227, New York, NY, USA, 2009. ACM
- Arijit Khan, Nan Li, Xifeng Yan, Ziyu Guan, Supriyo Chakraborty, and Shu Tao. Neighborhood based fast graph search in large networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 901–912, New York, NY, USA, 2011. ACM
- Juan P. Cedeño and K. Selçuk Candan. R2DF framework for ranked path queries over weighted RDF graphs. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 40:1–40:12, New York, NY, USA, 2011. ACM

.

Referências dos artigos subcategoria Integração de dados

- Jesse Weaver, Vito Giovanni Castellana, Alessandro Morari, Antonino Tumeo, Sumit Purohit, Alan Chappell, David Haglin, Oreste Villa, Sutanay Choudhury, Karen Schuchardt, and John Feo. Toward a data scalable solution for facilitating discovery of science resources. *Parallel Computing*, 40(10):682 – 696, 2014
- Renaud Delbru, Stephane Campinas, and Giovanni Tummarello. Searching web data: An entity retrieval and high-performance indexing model. *Web Semantics: Science, Services and Agents on the World Wide Web*, 10(0):33 – 58, 2012. Web-Scale Semantic Information Processing
- Thanh Tran, Haofen Wang, and Peter Haase. Hermes: Data web search on a pay-as-you-go integration infrastructure. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):189 – 203, 2009. The Web of Data
- Paul Groth, Antonis Loizou, Alasdair J.G. Gray, Carole Goble, Lee Harland, and Steve Pettifer. Api-centric linked data integration: The open PHACTS discovery platform case study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2014
- G. Karvounarakis, A. Magganaraki, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, and K. Tolle. Querying the semantic web with RQL. *Computer Networks*, 42(5):617 – 640, 2003. The Semantic Web: an evolution for a revolution
- A. Petermann, M. Junghanns, R. Muller, and E. Rahm. BIIIG: Enabling business intelligence with integrated instance graphs. In *IEEE 30th International Conference on Data Engineering Workshops (ICDEW)*, pages 4–11, March 2014
- J. Heer and A. Perer. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 51–60, Oct 2011
- W.E. Moustafa, A. Kimmig, A. Deshpande, and L. Getoor. Subgraph pattern matching over uncertain graphs with identity linkage uncertainty. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 904–915, March 2014
- Yunkai Liu and T.M. Vitolo. Graph data warehouse: Steps to integrating graph databases into the traditional conceptual structure of a data warehouse. In *IEEE International Congress on Big Data (BigData Congress)*, pages 433–434, June 2013
- Federica Mandreoli, Riccardo Martoglia, Giorgio Villani, and Wilma Penzo. Flexible query answering on graph-modeled data. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 216–227, New York, NY, USA, 2009. ACM
- María Constanza Pabón, Claudia Roncancio, and Martha Millán. Graph data transformations and querying. In *Proceedings of the 2014 International C* Conference on Computer Science & Software Engineering*, C3S2E '14, pages 20:1–20:6, New York, NY, USA, 2014. ACM

.

Referências dos artigos subcategoria Otimização de desempenho

- Jiwon Seo, S. Guo, and M.S. Lam. SociaLite: Datalog extensions for efficient social network analysis. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 278–289, April 2013

- Jun Gao, Jiashuai Zhou, Chang Zhou, and J.X. Yu. GLog: A high level graph analysis system using mapreduce. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 544–555, March 2014
- V. Dritsou, P. Constantopoulos, A. Deligiannakis, and Y. Kotidis. Shortcut selection in RDF databases. In *IEEE 27th International Conference on Data Engineering Workshops (ICDEW)*, pages 194–199, April 2011
- Letao Qi, H.T. Lin, and V. Honavar. Clustering remote RDF data using SPARQL update queries. In *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 236–242, April 2013
- G. Kasneci, M. Ramanath, M. Sozio, F.M. Suchanek, and G. Weikum. STAR: Steiner-tree approximation in relationship graphs. In *IEEE 25th International Conference on Data Engineering - ICDE'09*, pages 868–879, March 2009
- M. Jarrar and M.D. Dikaiakos. Querying the data web: The MashQL approach. *Internet Computing, IEEE*, 14(3):58–67, May 2010
- Fan Wang and Gagan Agrawal. Answering complex structured queries over the deep web. In *Proceedings of the 15th Symposium on International Database Engineering & Applications*, IDEAS '11, pages 115–123, New York, NY, USA, 2011. ACM
- Mirjana Mazuran, Edoardo Serra, and Carlo Zaniolo. Extending the power of datalog recursion. *The VLDB Journal*, 22(4):471–493, August 2013
- Semih Salihoglu and Jennifer Widom. GPS: A graph processing system. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, SSDBM, pages 22:1–22:12, New York, NY, USA, 2013. ACM
- Mohamed Sarwat, Sameh Elnikety, Yuxiong He, and Mohamed F. Mokbel. Horton+: A distributed system for processing declarative reachability queries over partitioned graphs. *Proc. VLDB Endow.*, 6(14):1918–1929, September 2013
- Juan P. Cedeño and K. Selçuk Candan. R2DF framework for ranked path queries over weighted RDF graphs. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 40:1–40:12, New York, NY, USA, 2011. ACM
- Jingren Zhou, Nicolas Bruno, Ming-Chuan Wu, Per-Ake Larson, Ronnie Chaiken, and Darren Shakib. SCOPE: Parallel databases meet MapReduce. *The VLDB Journal*, 21(5):611–636, October 2012

.

Tabela 6: Categoria: paradigma tecnológico.

Referências dos artigos subcategoria RDF
- Jesse Weaver, Vito Giovanni Castellana, Alessandro Morari, Antonino Tumeo, Sumit Purohit, Alan Chappell, David Haglin, Oreste Villa, Sutanay Choudhury, Karen Schuchardt, and John Feo. Toward a data scalable solution for facilitating discovery of science resources. *Parallel Computing*, 40(10):682 – 696, 2014
- Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker. Searching and browsing linked data with SWSE: The semantic web search engine. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):365 – 401, 2011. {JWS} special issue on Semantic Search

- Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of linked data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14(0):14 – 44, 2012. Special Issue on Dealing with the Messiness of the Web of Data
- Klara Weiand, Fabian KneiSSl, Wojciech Lobacz, Tim Furche, and François Bry. pest: Fast approximate keyword search in semantic data using eigenvector-based term propagation. *Information Systems*, 37(4):372 – 390, 2012. Semantic Web Data Management
- Xiang Lian, Eugenio De Hoyos, Artem Chebotko, Bin Fu, and Christine Reilly. k-nearest keyword search in RDF graphs. *Web Semantics: Science, Services and Agents on the World Wide Web*, pages 40 – 56, 2013
- Paul Groth, Antonis Loizou, Alasdair J.G. Gray, Carole Goble, Lee Harland, and Steve Pettifer. Api-centric linked data integration: The open PHACTS discovery platform case study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2014
- G. Karvounarakis, A. Magganaraki, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, and K. Tolle. Querying the semantic web with RQL. *Computer Networks*, 42(5):617 – 640, 2003. The Semantic Web: an evolution for a revolution
- A. Morari, V.G. Castellana, D. Haglin, J. Feo, J. Weaver, A. Tumeo, and O. Villa. Accelerating semantic graph databases on commodity clusters. In *IEEE International Conference on Big Data*, pages 768–772, Oct 2013
- V.G. Castellana, A. Tumeo, O. Villa, D. Haglin, and J. Feo. Composing data parallel code for a SPARQL graph engine. In *International Conference on Social Computing (SocialCom)*, pages 691–699, Sept 2013
- A. Cuzzocrea and P. Serafino. A reachability-based theoretical framework for modeling and querying complex probabilistic graph data. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1177–1184, Oct 2012
- Rui Wang and K. Chiu. A stream partitioning approach to processing large scale distributed graph datasets. In *IEEE International Conference on Big Data*, pages 537–542, Oct 2013
- O. Cure, F. Kerdjoudj, Chan Le Duc, M. Lamolle, and D. Faye. On the potential integration of an ontology-based data access approach in NoSQL stores. In *Third International Conference on Emerging Intelligent Data and Web Technologies (EIDWT)*, pages 166–173, Sept 2012
- M. Leida and A. Chu. Distributed SPARQL query answering over RDF data streams. In *IEEE International Congress on Big Data (BigData Congress)*, pages 369–378, June 2013
- P. Yadav and V. Samala. Benchmarking over a semantic repository. In *Second International Conference on Advanced Computing (ICoAC)*, pages 51–59, Dec 2010
- V. Dritsou, P. Constantopoulos, A. Deligiannakis, and Y. Kotidis. Shortcut selection in RDF databases. In *IEEE 27th International Conference on Data Engineering Workshops (ICDEW)*, pages 194–199, April 2011
- Letao Qi, H.T. Lin, and V. Honavar. Clustering remote RDF data using SPARQL update queries. In *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 236–242, April 2013
- M. Jarrar and M.D. Dikaiakos. Querying the data web: The MashQL approach. *Internet Computing, IEEE*, 14(3):58–67, May 2010

- Kai Zeng, Jiacheng Yang, Haixun Wang, Bin Shao, and Zhongyuan Wang. A distributed graph engine for web scale RDF data. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pages 265–276. VLDB Endowment, 2013
- Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 125–134, New York, NY, USA, 2012. ACM
- Lei Zou, M. Tamer Özsu, Lei Chen, Xuchuan Shen, Ruizhe Huang, and Dongyan Zhao. gStore: A graph-based SPARQL query engine. *The VLDB Journal*, 23(4):565–590, August 2014
- Amarnath Gupta and Simone Santini. On querying OBO ontologies using a DAG pattern query language. In *Proceedings of the Third International Conference on Data Integration in the Life Sciences*, DILS'06, pages 152–167, Berlin, Heidelberg, 2006. Springer-Verlag
- Juan P. Cedeño and K. Selçuk Candan. R2DF framework for ranked path queries over weighted RDF graphs. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 40:1–40:12, New York, NY, USA, 2011. ACM
- Padmashree Ravindra, HyeongSik Kim, and Kemafor Anyanwu. To nest or not to nest, when and how much: Representing intermediate results of graph pattern queries in mapreduce based processing. In *Proceedings of the 4th International Workshop on Semantic Web Information Management*, SWIM '12, pages 5:1–5:8, New York, NY, USA, 2012. ACM

.

Referências dos artigos subcategoria Paralelismo

- Jesse Weaver, Vito Giovanni Castellana, Alessandro Morari, Antonino Tumeo, Sumit Purohit, Alan Chappell, David Haglin, Oreste Villa, Sutanay Choudhury, Karen Schuchardt, and John Feo. Toward a data scalable solution for facilitating discovery of science resources. *Parallel Computing*, 40(10):682 – 696, 2014
- A. Morari, V.G. Castellana, D. Haglin, J. Feo, J. Weaver, A. Tumeo, and O. Villa. Accelerating semantic graph databases on commodity clusters. In *IEEE International Conference on Big Data*, pages 768–772, Oct 2013
- V.G. Castellana, A. Tumeo, O. Villa, D. Haglin, and J. Feo. Composing data parallel code for a SPARQL graph engine. In *International Conference on Social Computing (SocialCom)*, pages 691–699, Sept 2013
- Jun Gao, Jiashuai Zhou, Chang Zhou, and J.X. Yu. GLog: A high level graph analysis system using mapreduce. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 544–555, March 2014
- L. Fegaras. Supporting Bulk Synchronous Parallelism in Map-Reduce queries. In *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*, pages 1068–1077, Nov 2012
- Letao Qi, H.T. Lin, and V. Honavar. Clustering remote RDF data using SPARQL update queries. In *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 236–242, April 2013
- Zhao Sun, Hongzhi Wang, Haixun Wang, Bin Shao, and Jianzhong Li. Efficient subgraph matching on billion node graphs. *Proc. VLDB Endow.*, 5(9):788–799, May 2012

- Mohamed Sarwat, Sameh Elnikety, Yuxiong He, and Mohamed F. Mokbel. Horton+: A distributed system for processing declarative reachability queries over partitioned graphs. *Proc. VLDB Endow.*, 6(14):1918–1929, September 2013
- Jingren Zhou, Nicolas Bruno, Ming-Chuan Wu, Per-Ake Larson, Ronnie Chaiken, and Darren Shakib. SCOPE: Parallel databases meet MapReduce. *The VLDB Journal*, 21(5):611–636, October 2012
- Padmashree Ravindra, HyeongSik Kim, and Kemafor Anyanwu. To nest or not to nest, when and how much: Representing intermediate results of graph pattern queries in mapreduce based processing. In *Proceedings of the 4th International Workshop on Semantic Web Information Management*, SWIM '12, pages 5:1–5:8, New York, NY, USA, 2012. ACM
- Alan Chappell, Sutanay Choudhury, John Feo, David Haglin, Alessandro Morari, Sumit Purohit, Karen Schuchardt, Antonino Tumeo, Jesse Weaver, and Oreste Villa. Toward a data scalable solution for facilitating discovery of scientific data resources. In *Proceedings of the 2013 International Workshop on Data-Intensive Scalable Computing Systems*, DISCS-2013, pages 55–60, New York, NY, USA, 2013. ACM
- Shengqi Yang, Xifeng Yan, Bo Zong, and Arijit Khan. Towards effective partition management for large graphs. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 517–528, New York, NY, USA, 2012. ACM
.

Referências dos artigos subcategoria Distribuição de dados

- Yijun Bei, Zhen Lin, Chen Zhao, and Xiaojun Zhu. HBase system-based distributed framework for searching large graph databases. In *14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 151–156, July 2013
- M. Brocheler, A. Pugliese, and V.S. Subrahmanian. COSI: Cloud oriented subgraph identification in massive social networks. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 248–255, Aug 2010
- Rui Wang and K. Chiu. A stream partitioning approach to processing large scale distributed graph datasets. In *IEEE International Conference on Big Data*, pages 537–542, Oct 2013
- S.R. Spillane, J. Birnbaum, D. Bokser, D. Kemp, A. Labouseur, P.W. Olsen, J. Vijayan, Jeong-Hyon Hwang, and Jun-Weon Yoon. A demonstration of the G* graph database system. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 1356–1359, April 2013
- M. Leida and A. Chu. Distributed SPARQL query answering over RDF data streams. In *IEEE International Congress on Big Data (BigData Congress)*, pages 369–378, June 2013
- Kai Zeng, Jiacheng Yang, Haixun Wang, Bin Shao, and Zhongyuan Wang. A distributed graph engine for web scale RDF data. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pages 265–276. VLDB Endowment, 2013
- Dan Suciu. Distributed query evaluation on semistructured data. *ACM Trans. Database Syst.*, 27(1):1–62, March 2002
- Semih Salihoglu and Jennifer Widom. GPS: A graph processing system. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, SSDBM, pages 22:1–22:12, New York, NY, USA, 2013. ACM
.

Referências dos artigos subcategoria Linguagem declarativa

- Sherif Sakr, Sameh Elnikety, and Yuxiong He. Hybrid query execution engine for large attributed graphs. *Information Systems*, 41(0):45 – 73, 2014
- Jesse Weaver, Vito Giovanni Castellana, Alessandro Morari, Antonino Tumeo, Sumit Purohit, Alan Chappell, David Haglin, Oreste Villa, Sutanay Choudhury, Karen Schuchardt, and John Feo. Toward a data scalable solution for facilitating discovery of science resources. *Parallel Computing*, 40(10):682 – 696, 2014
- Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker. Searching and browsing linked data with SWSE: The semantic web search engine. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):365 – 401, 2011. {JWS} special issue on Semantic Search
- Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of linked data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14(0):14 – 44, 2012. Special Issue on Dealing with the Messiness of the Web of Data
- Timos K. Sellis, Nick Roussopoulos, and Raymond T. Ng. Efficient compilation of large rule bases using logical access paths. *Information Systems*, 15(1):73 – 84, 1990. Knowledge Engineering
- Klara Weiand, Fabian KneiSSl, Wojciech Lobacz, Tim Furche, and François Bry. pest: Fast approximate keyword search in semantic data using eigenvector-based term propagation. *Information Systems*, 37(4):372 – 390, 2012. Semantic Web Data Management
- Xiang Lian, Eugenio De Hoyos, Artem Chebotko, Bin Fu, and Christine Reilly. k-nearest keyword search in RDF graphs. *Web Semantics: Science, Services and Agents on the World Wide Web*, pages 40 – 56, 2013
- Paul Groth, Antonis Loizou, Alasdair J.G. Gray, Carole Goble, Lee Harland, and Steve Pettifer. Api-centric linked data integration: The open PHACTS discovery platform case study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2014
- G. Karvounarakis, A. Magganaraki, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, and K. Tolle. Querying the semantic web with RQL. *Computer Networks*, 42(5):617 – 640, 2003. The Semantic Web: an evolution for a revolution
- H.M. Jamil. Design of declarative graph query languages: On the choice between value, pattern and object based representations for graphs. In *IEEE 28th International Conference on Data Engineering Workshops (ICDEW)*, pages 178–185, April 2012
- R. Ronen and O. Shmueli. SoQL: A language for querying and creating data in social networks. In *ICDE'09. IEEE 25th International Conference on Data Engineering*, pages 1595–1602, March 2009
- A. Morari, V.G. Castellana, D. Haglin, J. Feo, J. Weaver, A. Tumeo, and O. Villa. Accelerating semantic graph databases on commodity clusters. In *IEEE International Conference on Big Data*, pages 768–772, Oct 2013
- J. Heer and A. Perer. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 51–60, Oct 2011

- V.G. Castellana, A. Tumeo, O. Villa, D. Haglin, and J. Feo. Composing data parallel code for a SPARQL graph engine. In *International Conference on Social Computing (SocialCom)*, pages 691–699, Sept 2013
- Jiwon Seo, S. Guo, and M.S. Lam. SociaLite: Datalog extensions for efficient social network analysis. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 278–289, April 2013
- A. Cuzzocrea and P. Serafino. A reachability-based theoretical framework for modeling and querying complex probabilistic graph data. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1177–1184, Oct 2012
- Rui Wang and K. Chiu. A stream partitioning approach to processing large scale distributed graph datasets. In *IEEE International Conference on Big Data*, pages 537–542, Oct 2013
- C. Raguenaud, J. Kennedy, and P.J. Barclay. The Prometheus taxonomic database. In *Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pages 63–70, 2000
- Jun Gao, Jiashuai Zhou, Chang Zhou, and J.X. Yu. GLog: A high level graph analysis system using mapreduce. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 544–555, March 2014
- O. Cure, F. Kerdjoudj, Chan Le Duc, M. Lamolle, and D. Faye. On the potential integration of an ontology-based data access approach in NoSQL stores. In *Third International Conference on Emerging Intelligent Data and Web Technologies (EIDWT)*, pages 166–173, Sept 2012
- M. Leida and A. Chu. Distributed SPARQL query answering over RDF data streams. In *IEEE International Congress on Big Data (BigData Congress)*, pages 369–378, June 2013
- P. Yadav and V. Samala. Benchmarking over a semantic repository. In *Second International Conference on Advanced Computing (ICoAC)*, pages 51–59, Dec 2010
- Taekyong Lee, Lei Sheng, T. Bozkaya, N.H. Balkir, Z. Meral Ozsoyoglu, and Z.M. Ozsoyoglu. Querying multimedia presentations based on content. *IEEE Transactions on Knowledge and Data Engineering*, 11(3):361–385, May 1999
- S. Dar and R. Agrawal. Extending SQL with generalized transitive closure. *IEEE Transactions on Knowledge and Data Engineering*, 5(5):799–812, Oct 1993
- V. Dritsou, P. Constantopoulos, A. Deligiannakis, and Y. Kotidis. Shortcut selection in RDF databases. In *IEEE 27th International Conference on Data Engineering Workshops (ICDEW)*, pages 194–199, April 2011
- Letao Qi, H.T. Lin, and V. Honavar. Clustering remote RDF data using SPARQL update queries. In *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 236–242, April 2013
- M. Jarrar and M.D. Dikaiakos. Querying the data web: The MashQL approach. *Internet Computing, IEEE*, 14(3):58–67, May 2010
- Kai Zeng, Jiacheng Yang, Haixun Wang, Bin Shao, and Zhongyuan Wang. A distributed graph engine for web scale RDF data. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pages 265–276. VLDB Endowment, 2013
- Fan Wang and Gagan Agrawal. Answering complex structured queries over the deep web. In *Proceedings of the 15th Symposium on International Database Engineering & Applications*, IDEAS '11, pages 115–123, New York, NY, USA, 2011. ACM

- Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 125–134, New York, NY, USA, 2012. ACM
- M. P. Consens and A. O. Mendelzon. Expressing structural hypertext queries in graphlog. In *Proceedings of the Second Annual ACM Conference on Hypertext*, HYPERTEXT '89, pages 269–292, New York, NY, USA, 1989. ACM
- Mirjana Mazuran, Edoardo Serra, and Carlo Zaniolo. Extending the power of datalog recursion. *The VLDB Journal*, 22(4):471–493, August 2013
- Lei Zou, M. Tamer Özsu, Lei Chen, Xuchuan Shen, Ruizhe Huang, and Dongyan Zhao. gStore: A graph-based SPARQL query engine. *The VLDB Journal*, 23(4):565–590, August 2014
- Mohamed Sarwat, Sameh Elnikety, Yuxiong He, and Mohamed F. Mokbel. Horton+: A distributed system for processing declarative reachability queries over partitioned graphs. *Proc. VLDB Endow.*, 6(14):1918–1929, September 2013
- Saumen Dey, Víctor Cuevas-Vicenttín, Sven Köhler, Eric Gribkoff, Michael Wang, and Bertram Ludäscher. On implementing provenance-aware regular path queries with relational query engines. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 214–223, New York, NY, USA, 2013. ACM
- Amarnath Gupta and Simone Santini. On querying OBO ontologies using a DAG pattern query language. In *Proceedings of the Third International Conference on Data Integration in the Life Sciences*, DILS'06, pages 152–167, Berlin, Heidelberg, 2006. Springer-Verlag
- Juan P. Cedeño and K. Selçuk Candan. R2DF framework for ranked path queries over weighted RDF graphs. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 40:1–40:12, New York, NY, USA, 2011. ACM
- Jingren Zhou, Nicolas Bruno, Ming-Chuan Wu, Per-Ake Larson, Ronnie Chaiken, and Darren Shakib. SCOPE: Parallel databases meet MapReduce. *The VLDB Journal*, 21(5):611–636, October 2012
- Padmashree Ravindra, HyeongSik Kim, and Kemafor Anyanwu. To nest or not to nest, when and how much: Representing intermediate results of graph pattern queries in mapreduce based processing. In *Proceedings of the 4th International Workshop on Semantic Web Information Management*, SWIM '12, pages 5:1–5:8, New York, NY, USA, 2012. ACM

Mineração em Grandes Massas de Dados Utilizando *Hadoop MapReduce* e Algoritmos Bio-inspirados: Uma Revisão Sistemática

Mining of Massive Data Bases Using Hadoop MapReduce and Bio-inspired algorithms: A Systematic Review

Sandro Roberto Loiola de Menezes [1]
Rebeca Schroeder Freitas [2]
Rafael Stubs Parpinelli [3]

Data de submissão: 24/08/2015, *Data de aceite: 06/02/2016*

Resumo: A Área de Mineração de Dados tem sido utilizada em diversas áreas de aplicação e visa extrair conhecimento através da análise de dados. Nas últimas décadas, inúmeras bases de dados estão tendenciando a possuir grande volume, alta velocidade de crescimento e grande variedade. Esse fenômeno é conhecido como *Big Data* e corresponde a novos desafios para tecnologias clássicas como Sistema de Gestão de Banco de Dados Relacional pois não tem oferecido desempenho satisfatório e escalabilidade para aplicações do tipo *Big Data*. Ao contrário dessas tecnologias, *Hadoop MapReduce* é um *framework* que, além de provêr processamento paralelo, também fornece tolerância a falhas e fácil escalabilidade sobre um sistema de armazenamento distribuído compatível com cenário *Big Data*. Uma das técnicas que vem sendo utilizada no contexto *Big Data* são algoritmos bio-inspirados. Esses algoritmos são boas opções de solução em problemas complexos multidimensionais, multiobjetivos e de grande escala. A combinação de sistemas baseados em *Hadoop MapReduce* e algoritmos bio-inspirados tem se mostrado vantajoso em aplicações *Big Data*. Esse artigo apresenta uma revisão sistemática de trabalhos nesse contexto, visando analisar critérios como: tarefas de mineração de dados abordadas, algoritmos bio-inspirados utilizados, disponibilidade das bases utilizadas e quais características *Big Data* são tratadas nos trabalhos. Como resultado, esse artigo discute os critérios analisados e identifica alguns modelos de paralelização, além de sugerir uma direção para trabalhos futuros.

[1]Programa de Pós-Graduação em Computação Aplicada, Centro de Ciências Tecnológicas, Universidade do Estado de Santa Catarina, Caixa Postal 631 - Joinville, Santa Catarina, Brasil. {sandrolmenezes@gmail.com}
[2]Departamento de Ciência da Computação, Centro de Ciências Tecnológicas, Universidade do Estado de Santa Catarina, Caixa Postal 631 - Joinville, Santa Catarina, Brasil. {rebeca.schroeder@udesc.br}
[3]Programa de Pós-Graduação em Computação Aplicada, Centro de Ciências Tecnológicas, Universidade do Estado de Santa Catarina, Caixa Postal 631 - Joinville, Santa Catarina, Brasil. {rafael.parpinelli@udesc.br}

***P*alavras-chave:** algoritmos bio-inspirados, Hadoop MapReduce, mineração de dados, big gata, revisão sistemática

Abstract: The Data Mining research field has been applied in several distinct areas and, in all of them, it aims to extract knowledge through data analysis. In recent decades, databases are tending to have large volume, high-speed growth rate and variety. This phenomenon is known as Big Data and represents new challenges for classical technologies such as Relational Database Systems. Unlike classical technologies, Hadoop MapReduce is a framework that, in addition to provide mechanisms for parallel processing, also offers fault tolerance and easy scalability over a distributed storage system compatible with Big Data. Some techniques that have been used in the context of Big Data are bio-inspired algorithms. These algorithms have being proved to be good solution options for complex multi-dimensional, multi-objectives and large scale problems. The mixing of Hadoop MapReduce and bio-inspired algorithms has proven to be advantageous in Big Data applications. This article presents a systematic review of Big Data applications in order to analyze: Which data mining tasks are being applied? What are the bio-inspired algorithms involved? Are the databases available to downlaod? and, What Big Data features are discussed in the works reviewed? As a result, this article discusses, analyzes and identifies some models of parallelization and suggests some directions for future research.

***K*eywords:** bio-inspired algorithms, Hadoop MapReduce, data mining, big data, systematic review

1 Introdução

Técnicas de mineração de dados têm recebido atenção por parte da comunidade científica devido aos benefícios que podem produzir em diversas áreas como Educação, Bioinformática, Economia, Negócios, entre outras. Tais benefícios podem ser obtidos através da descoberta de padrões, correlações e tendências de informações [47]. Esses padrões descobertos agregam conhecimento que podem ser úteis na tomada de decisão em um determinado processo. Na área de Negócios, por exemplo, dados de compras de clientes podem ser utilizados como fonte de mineração para obter um padrão de compras e assim embasar o planejamento estratégico de mercado [13]. Na área de segurança de redes, mineração de dados pode ser utilizada para auxiliar na detecção de invasão de intrusos [6]. No ramo das Ciências Biológicas, a mineração de dados contribui na descoberta de novos tipos de câncer [9]. Na Bioinformática, mineração de dados ajuda na identificação de espécies através da análise de DNA [14]. Na educação, mineração de dados tem descoberto novas técnicas de aprendizado e de avaliação de estudantes [47]. Além dessas áreas, mineração de dados vem

sendo introduzida em áreas como Análise de Dados Espaciais, Padrões Recognitivos, dentre outras [47].

Entretanto, existem alguns contextos de aplicação em que o processo de mineração de dados fica impossibilitado ou se torna inviável devido a existência de algum requisito que não é atendido de forma eficiente pelas tecnologias convencionais. Por exemplo, no contexto de grandes volumes de dados, um processo de mineração pode ocasionar em alto consumo de tempo de processamento. Outro fator limitante é a dimensionalidade dos dados que influencia diretamente no desempenho do algoritmo: quanto maior a dimensionalidade de dados, maior será o tempo para processa-los [28]. Além disso, as técnicas convencionais de mineração de dados precisam realizar a carga dos dados que serão analisados na memória antes de realizar a mineração [35]. No contexto de grandes volumes de dados, realizar esse carregamento na memória é mais um ponto crítico, pois delimita a capacidade máxima de dados que serão minerados em um determinado momento na estação de processamento [35].

Percebe-se que existem alguns requisitos especificamente relacionados com grandes volumes de dados, como disponibilidade e desempenho eficiente, que tecnologias convencionais de mineração de dados não são adequadas para cumpri-las. Por isso é necessário desenvolver tecnologias inovadoras que possam cumprir de forma eficiente esses requisitos. Para alcançar esse avanço tecnológico é necessário intensificar pesquisas científicas no contexto de mineração de grandes volumes de dados. O grande aumento do volume de dados que vem ocorrendo nas duas últimas décadas, fomentado por várias fontes de dados como mídias sociais, sensores, arquivo de *logs* dentre outras, reforça a necessidade de mineração de dados em grandes massas de dados.

Volume de dados na ordem de *Petabytes*, que corresponde a 10^{15} *Bytes*, é uma das propriedades de *Big Data* [11]. Segundo a definição de [1], *Big Data* é definido como grande volume de informações ativas com alta velocidade e variedade que demanda custo e formas inovadoras para obter uma visão compreensível para a tomada de decisão . Complementarmente, *Big Data* é considerado a próxima fronteira da inovação, competição e produtividade[11]. Além de um grande Volume de dados, *Big Data* também possui as propriedades de Velocidade, Variedade, Veracidade e Valor. Essas propriedades de *Big Data* são conhecidas como 5 V's e serão detalhadas no corpo deste artigo. Para realizar mineração de dados em bases com propriedades *Big Data*, principalmente Volume, é necessário que o processamento seja realizado em uma infraestrutura de processamento escalável, para que seja possível realizar a alocação de recursos de acordo com a demanda requerida pelo volume de dados a ser processado. A escalabilidade pode ser vertical ou horizontal.

A escalabilidade vertical é obtida através do incremento de mais recursos computacionais (núcleos, memória, processamento) em um servidor original, incrementando assim o poder de processamento [56][38]. Já escalabilidade horizontal é obtida através do incremento de um grande número de servidores menores e mais baratos, fortemente interconectados em

formato de *cluster* [56][38]. A escolha desse tipo de escalabilidade vem crescendo pois é mais acessível e possui melhor desempenho de processamento em relação à escalabilidade vertical [56]. Assim, a linha de pesquisa deste trabalho considera o contexto de escalabilidade horizontal. Utilizando escalabilidade horizontal é possível realizar processamento distribuído e paralelo. Diante disso, nesse contexto é necessário utilizar técnicas com características mais adequadas não só ao processamento em paralelo como também ao cenário *Big Data*. Algoritmos bio-inspirados possuem conceitos relacionados a cooperação e coevolução que são vantajosas no contexto de paralelismo [44]. Além disso, esses algoritmos possuem algumas características que favorem a sua escolha para trabalhar com problemas de grande escala.

Algoritmos Bio-inspirados pertencem a uma subárea da Computação conhecida como Computação Natural, que utiliza a natureza como fonte de inspiração para desenvolver técnicas computacionais não-determinísticas para resolver problemas complexos de otimização [21]. Esses algoritmos englobam abordagens populacionais que processam um conjunto de indivíduos no qual cada indivíduo representa uma possível solução para o problema. Essa característica populacional tem contribuído na resolução de problemas de grande escala, alta dimensionalidade e multiobjetivos [12]. Outra característica vantajosa desses paradigmas bio-inspirados é o não determinismo aliado com a intensificação e diversificação. Isso implica em uma exploração mais eficiente do espaço de busca e em diferentes regiões [46]. Mineração em grandes massas de dados envolve espaço de busca de solução bastante extenso. Nesse caso, a diversificação pode beneficiar a busca da melhor solução. Soluções determinísticas tendem a falhar quando aplicadas em problemas que o espaço de busca é muito grande e tendem a aumentar consideravelmente [10] [54]. Mineração de dados em *Big Data* pode ser visto como um problema de otimização onde busca-se encontrar a melhor solução, como os valores dos centróides de uma tarefa de agrupamento, em um determinado tempo limite. Algoritmos Bio-inspirados são técnicas de pesquisa e otimização que possuem características adequadas para análise de dados com propriedades *Big Data* [12].

No contexto *Big Data*, é necessário embasar a técnica escolhida para realizar a mineração de dados numa infraestrutura que dê suporte de armazenamento e processamento para o desenvolvimento do mecanismo para mineração em grandes volumes de dados. Ferramentas convencionais como Sistemas de Gerenciamento de Banco de Dados Relacional não conseguem gerenciar *Big Data* com disponibilidade e desempenho eficientes [48] [19]. *Hadoop* é um projeto *open source* criado pela *Apache Software Fundation* que vem sendo utilizado como ferramenta no contexto *Big Data* [7]. A utilização dessa ferramenta no contexto *Big Data* é justificada pelo fornecimento de um *framework* de processamento em paralelo com tolerância a falhas, fácil escalabilidade e utilização (*Hadoop MapReduce*) [34] [52], além de um poderoso sistema de arquivos robusto e distribuído (*Hadoop Distributed File System (HDFS)*).

Existem outras tecnologias disponíveis que fornecem paralelismo que poderiam ser

utilizadas neste trabalho como *Graphical Processing Unit (GPU)*, *Grid Computing* e *Message Passing Interface (MPI)*. *GPU* é uma tecnologia orientada a transferência de dados, *throughput-oriented*. Essa tecnologia adapta-se bem aos problemas que envolvem transferência e carga de dados, *workload* [42]. *Grid Computing* é uma tecnologia que disponibiliza um conjunto de recursos, como estações de trabalho e rede com alta velocidade, para serem utilizadas sob demanda à medida que as tarefas de processamento são executadas. Essa tecnologia é recomendada na resolução de problema computacional de grande escala, pois provê grande desempenho [2]. *MPI* é uma interface eficiente e flexível que provê um padrão de comunicação seguro, veloz e de alto desempenho para aplicações distribuídas em *cluster* [37]. Aplicações *MPI* fornecem tolerância a falhas utilizando pontos de checagem, *checkpoints*, ocasionando o problema de excesso de armazenamento em disco [33]. Dentre essas tecnologias citadas, o *Hadoop* provê um mecanismo automático e mais eficiente de tolerância a falhas [39] [26] conciliado com escalabilidade e paralelismo de alto desempenho [53].

O objetivo desse artigo será alcançado através da revisão sistemática da literatura [29]. Essa revisão busca identificar quais algoritmos bio-inspirados são utilizados, como o algoritmo utiliza a infraestrutura fornecida pelo *Hadoop*, quais bases são utilizadas e, por fim, caracterizar os trabalhos encontrados em modelos de execução.

Esse artigo está estruturado da seguinte maneira. A Seção 2 explica os conceitos relacionados à mineração de dados e *Big Data*. A Seção 3 detalha Algoritmos Bio-Inspirados. Na Seção 4 é detalhada a estrutura do *Hadoop*. Na Seção 5 será apresentado o método de pesquisa científica adotado nesse artigo. Na Seção 6 será exposta a síntese dos trabalhos relacionados com mineração de dados utilizando algoritmos bio-inspirados e *Hadoop* . Na Seção 7 é exposta a discussão sobre os trabalhos revisados e na Seção 8 é feita a conclusão dessa pesquisa.

2 Mineração de dados e *Big Data*

Mineração de dados, também conhecido como Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases KDD*), é o processo de extração automática de padrões que representa conhecimento implícito armazenado ou capturado em grandes bases de dados, na *Web*, em fluxo de dados ou em outros repositórios de informações [24]. O processo de análise envolve técnicas matemáticas, estatísticas e computacionais. Processos ou tarefas no domínio de mineração de dados visam construir modelos eficientes capazes de analisar e extrair conhecimento além de predizer tendências futuras no comportamento dos dados. Tarefas de mineração de dados podem ser classificadas em duas categorias: preditiva e descritiva. Modelos preditivos aprendem através da análise de um conjunto de dados, ou dados de treinamento, e fazem predições no comportamento de novos dados. Já técnicas descritivas proveem sumarização de dados. Dentre as tarefas mais comuns de mineração de

dados é possível citar: Classificação, Agrupamento, Associação e Regressão [22]. As tarefas de Classificação e Agrupamento são detalhadas na sequência por serem abordadas nos trabalhos revisados.

A classificação é uma tarefa preditiva que infere a determinação de qual classe um novo dado está classificado. Basicamente, essa tarefa divide-se em duas fases. Na primeira fase ocorre o processo de aprendizagem. Nesta fase o modelo preditivo que define qual classe um determinado dado pertence é construído a partir da análise de um conjunto de dados, ou dados de treinamento. Na maioria das vezes os dados de treinamento são rotulados, ou seja, cada dado já possui a informação correta de qual classe ele pertence (atributo meta ou descritor alvo). Nesses casos a classificação é dita como supervisionada. Na segunda fase ocorre a avaliação do modelo em dados de teste, disjunto dos dados de treinamento, para encontrar a precisão do padrão de classificação [22] [30]. Aplicar essa tarefa em grandes massas de dados com número muito grande de atributos e com características redundantes pode trazer prejuízos a precisão de classificação [36]. Desta maneira, é comum antes de realizar a análise para gerar o modelo de classificação executar uma etapa de seleção de atributos que serão considerados pelo algoritmo de aprendizagem [23].

A tarefa de agrupamento tem como principal objetivo dividir um conjunto de objetos de dados não rotulados em diferentes grupos. Cada grupo é formado por objetos que possuem uma maior similaridade entre os mesmos. Para determinar se a técnica utilizada no processo de agrupamento é eficiente deve ser mensurada a qualidade dos grupos formados. Esta medida de qualidade é feita através do cálculo da similaridade entre os elementos de um mesmo grupo e de grupos diferentes. Quanto maior é a similaridade dos elementos de um mesmo grupo maior é a qualidade do agrupamento. Dessa forma a similaridade interna do grupo deve ser maximizada. Quanto menor for a similaridade de objetos de grupos diferentes, maior será a qualidade do agrupamento, logo essa similaridade deve ser minimizada [5].

Executar tarefas de Mineração de Dados no contexto *Big Data* é bastante dificultoso. Portanto, é importante entender o que significa *Big Data*. O termo *Big Data* pode ser definido como um cenário que envolve imensa quantidade de dados não estruturados que requer tecnologias não convencionais para analisar os mesmos. Além do volume e estrutura, outras características tornam *Big Data* um problema muito complexo para ser resolvido utilizando sistema de banco de dados convencional como Sistemas Gerenciadores de Banco de Dados Relacionais [28] [30] [48].

A necessidade de desenvolver novas tecnologias de processamento para *Big Data* surgiu em vários domínios devido ao aumento do poder computacional, capacidade de armazenamento e aumento na produção de conteúdo digital [17]. Pode-se citar como exemplo de domínio *Big Data* as áreas da Ciência, Telecomunicação, Indústria, Negócios, Planejamento urbano, Mídia social e Saúde. *Big Data* provê grande potencial no processo decisório baseado em dados podendo trazer benefícios como nova visão de negócio, habilidade de medir e

monitorar fatores influentes no negócio, descoberta de novas oportunidades de vendas dentre outros benefícios. Dessa forma, investimentos direcionados para pesquisa e desenvolvimento relacionados com *Big Data* vem ganhando espaço nas Universidades, governo e indústria. Apesar de existirem várias questões em discussão sobre *Big Data*, as propriedades ou características mais significativas do mesmo são identificadas a seguir [18]. Essas características são referenciadas pelo termo 5 V's: Volume, Velocidade, Variedade, Veracidade e Valor.

Em 2003 todo o volume de dados gerado e armazenado no mundo inteiro foi inferior a 1.8 *Zettabytes*. Em apenas dois dias no ano de 2011 o volume de dados gerado foi superior a 1.8 *Zettabytes*, conforme relatório fornecido pela *International Data Corporation (IDC)* [11]. A tendência é de aumentar o volume de bancos de dados corporativos em 40 % a cada ano. Entre outros fatores, os avanços e barateamento das tecnologias de coleta e armazenamento possibilitaram às aplicações computacionais fornecerem funcionalidades como operações de gravação e recuperação de dados com uma capacidade superior ao fornecido anteriormente. Além disso, o rápido desenvolvimento de redes e o aumento da capacidade de coleta de dados são outros fatores que influenciaram na tendência de aumento de volume de dados [51].

O Volume é a característica representada pela palavra *Big* de *Big Data*. Essa característica impõe requisitos específicos para tecnologias e ferramentas utilizadas atualmente para manipular *Big Data* [18]. No contexto *Big Data* atualmente os dados existentes estão em *Petabytes* e é previsível o aumento para *Zettabytes* num futuro próximo [28]. As redes sociais existentes estão produzindo dados na ordem de *Terabytes* todos os dias e essa quantidade de dados é difícil de ser manipulada usando sistemas tradicionais.

A Velocidade no contexto *Big Data* lida com a velocidade da chegada de dados de várias fontes e com a velocidade em que os dados fluem [28]. Nesse caso, os mesmos são frequentemente gerados em alta velocidade, como dados gerados por matrizes de sensores ou múltiplos eventos [18]. Além desses dados serem gerados com uma alta velocidade eles também precisam ser processados em tempo real ou próximo disso. Como exemplo dessa característica pode-se citar os dados produzidos por dispositivos de sensores que seriam constantemente enviados para armazenamento em banco de dados.

No intuito de obter informações baseadas no histórico de dados armazenados, aplicações de *software* podem executar certas consultas numa determinada base de dados e assim poder deduzir importantes resultados. Essas informações podem auxiliar os usuários a encontrar tendências de negócio dando a possibilidade de alterar as suas estratégias [28]. Assim, pode-se perceber que existe um grande valor contido nos dados armazenados e que pode levar a muitas vantagens para a indústria e comércio, dentre outros ramos. Segundo [18], o valor é uma importante característica de um dado que é definida pelo valor agregado que o dado coletado pode trazer para um processo, atividade ou hipótese.

A Veracidade considera a inconsistência no fluxo de dados. O carregamento de dados

torna-se um desafio de ser manutenido. Especialmente em redes sociais com o incremento no uso que geram picos de carregamento de dados com a ocorrência de certos eventos [28]. Segundo [18], a Veracidade no contexto *Big Data* inclui principalmente dois aspectos: consistência dos dados que podem ser definidas por sua confiabilidade estatística e a confiabilidade dos dados definida pelo número de fatores incluindo a origem dos dados, métodos de coleta, processamento e infraestrutura confiável. A veracidade *Big Data* garante que o dado usado é confiável, autêntico e protegido de acessos e modificações não autorizadas [18].

A Variedade de dados está relacionada com os tipos de dados que são gerados na formação do *Big Data*. Segundo [28], os dados produzidos não são apenas dados tradicionais ou estruturados, mas também semiestruturados e não-estruturados como: páginas *web*, arquivos de *log* , e-mails, documentos, dados de sensores de dispositivo, dados de redes sociais, dentre outros. A variedade impõe novos requisitos para armazenamento de dados e projetos de banco de dados que devem adaptar-se dinamicamente ao formato dos dados que estão sendo manipulados [18].

Diante da complexidade imposta ao processo de mineração de dados com as características 5 V's como, por exemplo, o processamento de grandes massas de dados com alta dimensionalidade considerando restrições relacionadas com recursos e tempo, torna-se necessário utilizar técnicas de otimização já consolidadas na resolução de problemas complexos do mundo real. A próxima seção descreve os principais algoritmos bio-inspirados encontrados na literatura e discute porque tais algoritmos podem ser utilizados como técnicas para manipular *Big Data*.

3 Algoritmos Bio-Inspirados

Algoritmos bio-inspirados tem sido utilizados como métodos de otimização para problemas complexos normalmente não estacionários e multi-dimensionais [49]. A natureza tem sido utilizada como fonte de inspiração para o desenvolvimento de algoritmos de otimização, sendo denominados de Algoritmos Bio-Inspirados. Segundo [21], essa categoria de algoritmos pode ser subdividida em quatro grupos. São eles: Computação Evolucionária, Inteligência de Enxame, Redes Neurais Artificiais e Sistemas Imunológicos Artificiais. Esse artigo aborda algoritmos pertencentes à Inteligência de Enxame e Computação Evolucionária visto que apenas algoritmos pertencentes a esses dois grupos foram utilizados nos artigos revisados. É importante mencionar que Inteligência de Enxame e Computação Evolucionária possuem a semelhança de representarem meta-heurística estocásticas e serem baseadas em população de possíveis soluções, aqui abstraídas pelo termo indivíduo.

Em uma população de indivíduos, cada indivíduo representa uma potencial solução do problema que está sendo otimizado. A população de indivíduos tende a mover-se para áreas onde estão localizadas as melhores soluções considerando todo o espaço de solução

de um determinado problema. Além disso, Inteligência de Enxame e Computação Evolucionária mantem e sucessivamente melhoram uma população de potenciais soluções até que alguma condição de parada seja satisfeita . A métrica utilizada para determinar o quanto uma solução é boa para um determinado problema é uma função de eficiência ou função *fitness*. Através dessa função é possível determinar qual é o melhor indivíduo da população em um determinado momento [12].

Inteligência de Enxame é um conjunto de técnicas de pesquisa e otimização que são inspiradas no comportamento social de alguns organismos vivos como formigas, abelhas, pássaros, baratas, dentre outros. Auto-organização e controle descentralizado são características essenciais da Inteligência de Enxame. Essas características conduzem ao comportamento emergente que surge através de interações locais entre componentes do sistema e não pode ser obtido por um componente atuando sozinho [43]. Inteligência de Enxame possui algumas meta-heurísticas inspirados na natureza. Otimização por Enxame de Partícula, Otimização por Colônia de Formigas e Otimização por Enxame de Vermes Luminescentes são exemplos de algoritmos pertencentes a Inteligência de Enxame que foram abordados nos artigos revisados.

A Computação Evolucionária é outro paradigma inserido na Computação Natural e baseia-se na teoria de Charles Darwin para simular o processo evolutivo. A teoria de Darwin apresenta o processo de seleção natural que ocorre na natureza evidenciando a sobrevivência dos mais bem adaptados [57]. No contexto computacional, cada indivíduo representa uma solução candidata do espaço de solução do problema tratado. Computação Evolucionária é composta por vários subprocessos inspirados na evolução das espécies. São eles: seleção, reprodução, combinação ou *crossover* e mutação. Algoritmos Genéticos, Algoritmo *CHC*, Evolução Diferencial e Programação Genética são exemplos de algoritmos evolucionários utilizados nos artigos revisados.

A seguir são detalhados alguns algoritmos pertencentes à Inteligência de Enxame e Computação Evolucionária. A relação entre esses algoritmos e as tarefas de mineração de dados será apresentada na Seção 6.

3.1 Otimização por Enxame de Partícula

Em 1995, Otimização por Enxame de Partícula (*Particle Swarm Optimization - PSO*) foi proposta por Kennedy [32]. Esse algoritmo é inspirado no comportamento coordenado dos pássaros ao voarem e do movimento de cardume de peixes. O peixe ou o pássaro é considerado como uma partícula no algoritmo. No contexto computacional, cada partícula representa uma solução do problema. A população de partículas é inicializada de forma aleatória dentro do domínio do espaço de solução do problema. Além da solução, cada partícula armazena uma posição no espaço de busca, uma velocidade de deslocamento no espaço de

busca e sua melhor posição até a iteração atual. No processamento do enxame de partículas, a melhor posição de todas as partículas também é armazenada, chamada de solução global. No momento da atualização da velocidade de uma partícula, tanto a melhor posição pessoal, representando a intensificação do algoritmo, quanto a melhor posição global do enxame, representando a diversificação, são consideradas de forma estocástica na atualização da velocidade. Após a atualização da velocidade também é atualizada a nova posição da partícula no espaço de solução para avaliar o quanto essa nova solução é boa. Com isso será possível avaliar se a solução é melhor que a melhor solução pessoal e global. Esse processo é executado até o critério de parada ser satisfeito. Número de gerações, número de avaliação de função *fitness* e estagnação da melhor solução podem ser utilizados como critérios de parada. Os parâmetros do *PSO* são: o tamanho da população, o peso da inércia e os coeficientes de aceleração.

3.2 Otimização por Colônia de Formigas

Proposto por Dorigo em 1999, a Otimização por Colônia de Formigas (*Ant Colony Optimization - ACO*) é inspirada no comportamento das formigas pela busca de alimentos [31]. Foi observado que, no decorrer do tempo, as formigas são capazes de descobrir o caminho mais curto entre a fonte de alimento e sua colônia, ou seja, o melhor caminho. Isso ocorre graças ao rastro de feromônio que a formiga deposita no caminho percorrido entre a fonte de alimento e a colônia. Como o feromônio é uma substância atrativa para as formigas, quando elas depositam esta substância no caminho, o mesmo fica mais atrativo e recruta outras formigas para este caminho. Entretanto, quando diminuir a quantidade de formigas em determinado caminho, o feromônio tende a diminuir pois essa substância é volátil e evapora rapidamente. No contexto computacional, o caminho da formiga representa uma solução para um problema. Os parâmetros definidos no *ACO* são: peso relativo da trilha, a atratividade da trilha, o tamanho da população e a taxa de evaporação do feromônio.

3.3 Otimização por Enxame de Vermes Luminescentes

Otimização por Enxame de Vermes Luminescentes (*Glowworm Swarm Optimization - GSO*) foi proposto por Krishnanand e Ghose em 2005. Foi inspirado na capacidade do controle de emissão de luminosidade própria desses vermes em determinadas situações. Por exemplo, para atrair a fêmea ou para atrair uma presa numa determinada área de abrangência de formato circular [32]. A substância que influencia no nível de luminosidade é a luciferina. O verme escolhe a direção para onde se movimentar baseado na intensidade de luminescência das regiões vizinhas. Quanto mais intensa a luciferina mais atrativa é a região. Com o tempo de permanência do indivíduo em regiões sem luminosidade o nível de luciferina do mesmo tende a diminuir. Computacionalmente, cada verme luminoso representa uma solução no espaço de solução do problema. A região que possui indivíduos com mais luciferina está mais

próxima da solução ótima. Dessa forma, baseando-se apenas em informações locais os indivíduos vão se movimentando até se compactarem numa determinada região. A função *fitness* é atualizada considerando o nível de luciferina encontrada na região definida por um raio além da taxa de decaimento da luciferina. O *GSO* possui os seguintes parâmetros: número de indivíduos, taxa de decaimento da luciferina, raio de abrangência da vizinhança e ponderador da função objetiva considerada no cálculo do nível de luciferina.

3.4 Algoritmos Genéticos

Propostos por John Holland em 1975, os Algoritmos Genéticos (*Genetic Algorithms - GA*) são os mais conhecidos e utilizados dentre os Algoritmos Evolucionários [15]. São inspirados no processo de Seleção Natural, segundo Darwin, onde os indivíduos mais bem adaptados ao meio possuem maior chance de sobreviver e assim reproduzir mais dependentes. Além da seleção natural, a hereditariedade genética também é inspiração desses algoritmos. Cada indivíduo, também chamado de cromossomo, possui um conjunto de tamanho fixo de *bits*, ou genes, que são representados em uma base de codificação. A reprodução desses indivíduos é realizada através da recombinação, ou *crossover*, de dois indivíduos pré-selecionados. Após isso, ainda na reprodução ocorre o processo de mutação nos descendentes recém-gerados. As características de intensificação e diversificação são expressos, respectivamente, através da combinação e mutação. Parâmetros do *GA* são: o tamanho da população, a probabilidade de *crossover* e probabilidade de mutação.

3.5 Evolução Diferencial

Evolução Diferencial (*Differential Evolution - DE*) foi proposta por Storn e Price [50]. Também possui uma população de indivíduos onde cada indivíduo possui um tamanho fixo e representa uma solução. O indivíduo da população é representado por um vetor e cada indivíduo possui um custo calculado com base no próprio vetor. Os indivíduos inicialmente terão os vetores inicializados com valores aleatórios dentro do domínio de solução do problema. A seleção de indivíduos para reprodução é feita de forma aleatória. Seleciona-se um indivíduo alvo da reprodução e mais três indivíduos aleatoriamente. O processo de reprodução ocorre através da adição da diferença ponderada entre dois dos três indivíduos ao terceiro indivíduo resultando num novo indivíduo. Nesta etapa ocorre a operação de mutação que caracteriza a diversificação. Após isso, ocorre o processo de combinação ou *crossover* do indivíduo mutado e o indivíduo alvo. Como resultado da combinação, resulta-se o indivíduo teste. A combinação caracteriza a intensificação do algoritmo *DE*. Após isso é avaliado o custo do indivíduo teste e do indivíduo alvo. Se o indivíduo teste possuir custo inferior ao do indivíduo alvo ele substituirá o indivíduo alvo na próxima geração. Caso contrário, o indivíduo teste será descartado. Serão produzidas novas gerações até que o critério de parada seja atendido. Normalmente o critério de parada é o número de gerações. Os parâmetros utilizados no *DE*

são: o tamanho da população, a probabilidade de *crossover* e o fator de mutação, número de dimensões do vetor, número de gerações.

3.6 Programação Genética

Em 1992 John Koza utilizou *GA's* para desenvolver programas computacionais aplicados em certas tarefas computacionais, surgindo assim a Programação Genética (*Genetic Programming - GP*) [31]. O indivíduo da população é um programa composto por funções e terminais. Os indivíduos são de tamanhos variados além de serem representados por árvores. As folhas das árvores são valoradas com variáveis e os nós internos com funções. A população inicial de indivíduos é gerada de forma aleatória considerando o domínio de solução do problema. Após avaliar o *fitness* de toda população de programas, são selecionados alguns programas para reprodução. A seleção pode ser feita com base em torneio ou roleta dentre outros critérios de seleção. Na reprodução de indivíduos para próxima geração além do *crossover* e mutação existe também o operador de reprodução. Na operação de *crossover* um ponto aleatório é escolhido para corte em cada árvore e as sub árvores geradas após o corte são trocadas. Na mutação uma folha ou nó de uma árvore é alterada de forma aleatória. Já na operação de reprodução o programa é copiado sem qualquer alteração para a próxima geração. Esse processo é realizado até que seja satisfeito algum critério de parada que normalmente considera o número de gerações. Nesse processo evolutivo os programas vão evoluindo de tal forma que o melhor indivíduo da última geração consiga representar uma função de aproximação do problema analisado. São parâmetros da *GP*: tamanho da população, probabilidade de *crossover*, probabilidade de mutação e número de gerações.

No cenário *Big Data*, para obter um melhor desempenho no processamento de algoritmos bio-inspirados é necessário embasar esses algoritmos numa infraestrutura com características adequadas a esse contexto e assim maximizar os benefícios dessa técnica. A ferramenta *Hadoop*, detalhado na Seção seguinte, fornece a infraestrutura com recursos mais adequados à técnica de mineração de dados.

4 *Hadoop*

Hadoop é uma ferramenta de código aberto utilizada para armazenar e manipular grandes massas de dados [41]. Com essa ferramenta o grande problema existente relacionado com tempo no processo de leitura de uma grande massa de dados é amenizado com a leitura de múltiplos discos de forma simultânea. Nesse tipo de leitura podem ocorrer algumas falhas que também são contornadas através da replicação de dados. Esse armazenamento seguro e compartilhado é provido pelo componente *Hadoop Distributed File System (HDFS)* [28]. Outro componente provido pelo *Hadoop* é o sistema de análise chamado de *framework*

MapReduce. *MapReduce* explora a arquitetura de armazenamento distribuída do *HDFS* provendo assim escalabilidade, confiabilidade e processamento paralelo [40]. Nas subseções a seguir, os componentes *HDFS* e *Hadoop MapReduce* serão detalhados.

4.1 *HDFS*

HDFS é um sistema de arquivos projetado para armazenar arquivos de dados muito grande com um padrão contínuo de acesso de dados e em *cluster*. *HDFS* possui arquitetura cliente-servidor e utiliza o protocolo de comunicação *Transmission Control Protocol (TCP)*. Assim, existem dois tipos de nodo. Um deles é o servidor, chamado *namenode* ou referenciado também por *Master*. O *Master* é responsável por gerenciar a localização de blocos de arquivos fragmentados e replicados do sistema de arquivos, além de manutenir os metadados e a árvore do sistema de arquivos. O outro tipo de nodo é o cliente, chamado de *datanode* ou também referenciado por *Worker*. O *Worker* possui como função armazenar e recuperar blocos de dados solicitados por aplicações de *software* ou pelo *Master*. Em alguns casos a utilização de *HDFS* pode não ser adequada, como nos casos que envolvem o contexto de baixa latência de acesso de dados porque o tempo requerido para inicialização do serviço, divisão e junção das tarefas, além da comunicação entre os nós certamente inviabiliza a utilização do mesmo [28].

4.2 *MapReduce*

MapReduce é um modelo de programação que possui principalmente duas funções, uma chamada de *map* e outra chamada de *reduce*. Cada uma dessas funções recebe como entrada de dados um conjunto de pares chave-valor e após o processamento dessas informações, de acordo com a função implementada pelo programador, tem como saída um conjunto de pares chave-valor. Basicamente existem duas fases de execução no processamento *MapReduce*, cada fase executando uma função. Na primeira fase é executado o processo de mapeamento através da função *map* e na segunda fase é executado o processo de redução através da função *reduce*. Essas fases podem ser executadas em paralelo através dos *datanodes* disponíveis em um determinado *cluster*. O tempo de processamento depende do volume de dados processados e o método de distribuição e replicação dos dados no *cluster*. Quanto maior o volume de dados maior será o tempo de processamento e quanto maior o *cluster* menor será o tempo de execução [40]. Um *job MapReduce* representa um processo completo de execução do *framework MapReduce* num determinado arquivo *HDFS*.

A Figura 1 mostra uma visão geral do fluxo de execução de um programa utilizando o *framework MapReduce*.

No fluxo com rótulo (1) a biblioteca *MapReduce* faz algumas cópias do programa do usuário para as estações *Master* e *Workers* do *cluster*. No fluxo com rótulo (2) o nodo *Master*

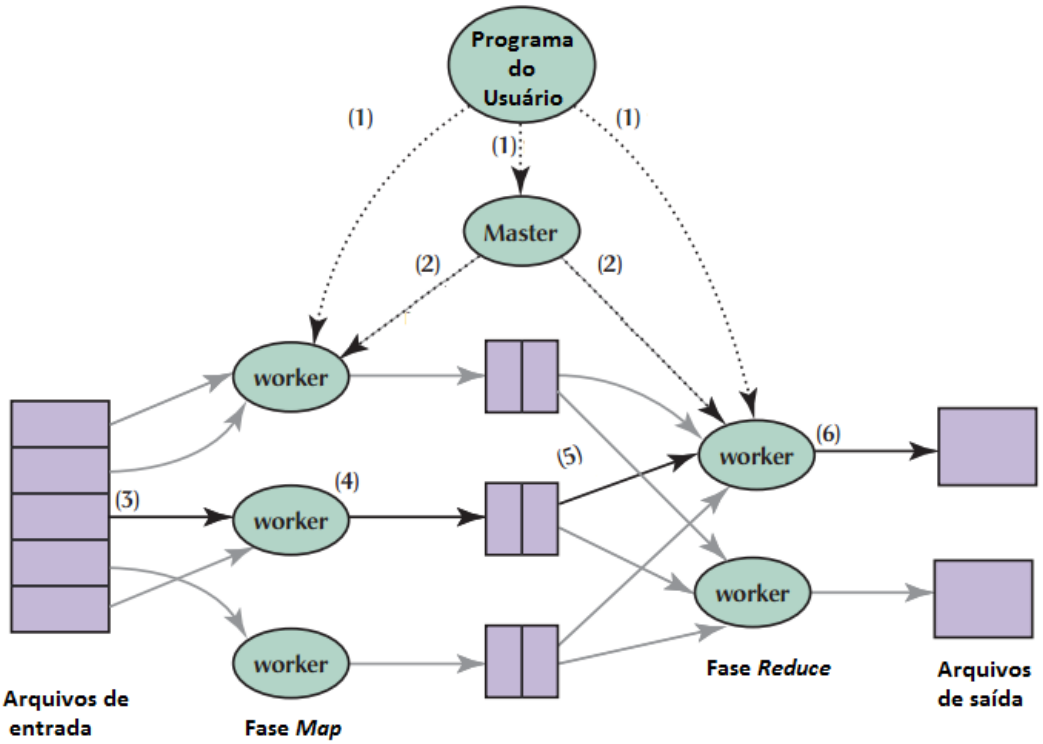


Figura 1. Execução *job MapReduce*. Adaptado de [16].

faz a atribuição das tarefas *map* e *reduce* para os *workers*. No fluxo cujo rótulo é (3) os *workers* que irão realizar a execução da função *map* leem o conteúdo correspondente ao dado de entrada e transformam esses dados em pares no formato chave-valor. Os dados nesse formato são passados como entrada para a função *map* que é implementada pelo programador. A saída de dados da função *map* também está em formato de pares chave-valor e fica armazenada na memória. No fluxo correspondente ao rótulo (4) os dados de saída da função *map* que estão em memória são gravados em disco local, chamados de arquivos intermediários, e posteriormente é enviada a localização desses arquivos para o *Master*. No fluxo (5) os *workers* responsáveis pra fazer a redução são notificados pelo *Master* com a informação de localização dos arquivos intermediários. Assim os *workers* fazem a leitura remota dos arquivos intermediários. Os *workers* de redução ordenam os dados no formato de pares chave-valor de tal forma que todos os valores de uma mesma chave são agrupados e relacionados àquela chave. Assim todos os pares terão chave diferente e cada chave terá o conjunto de valores relacionado a ela. Em seguida, ainda no fluxo de número (5) cada chave com seus valores são enviados para a função *reduce* processar de acordo com a implementação realizada pelo

programador. Após o processamento, representado pelo fluxo de número (6), a saída de dados é persistida em um arquivo final. Ao finalizar todas as tarefas de mapeamento e redução, o *Master* retorna para o programa do usuário [16].

Visto que Algoritmos Bio-inspirados embasados no *framework Hadoop MapReduce* podem compor boas soluções no processo de mineração de *Big Data*, esse artigo faz a revisão da literatura de trabalhos que atuam nesse contexto de solução utilizando o método descrito na próxima Seção.

5 Metodologia

A revisão sistemática da literatura proposta por [29] foi o método de pesquisa escolhido para a realização da pesquisa contida nesse artigo. Isso reduz a possibilidade de obter resultados tendenciosos além de produzir artefatos capazes de viabilizar a reprodução da revisão apresentada nesse artigo. As próximas seções detalham os principais componentes do protocolo utilizado.

A revisão sistemática exibida nesse artigo visa responder as seguintes questões de pesquisa.

- Pergunta Principal

Questão 1 - Quais são os *softwares* aplicados em mineração de grandes massas de dados que utilizam algoritmos bio-inspirados projetados em *Haddop MapReduce*?

- Perguntas Secundárias

Questão 2 - Em quais tarefas ou etapas de mineração de dados o mecanismo é aplicado?

Questão 3 - Quais são os algoritmos bio-inspirados utilizados?

Questão 4 - Quantos algoritmos bio-inspirados são paralelizados?

Questão 5 - Como o algoritmo está projetado no *Hadoop MapReduce*?

Questão 6 - Quais bases utilizadas estão disponíveis?

Questão 7 - Quais V's são considerados?

A estratégia de busca está dividida em duas partes: chave de busca e fonte de pesquisa.

A chave de busca representa a expressão que será utilizada pelos mecanismos de busca para realizar a pesquisa. Assim, a chave foi definida com a composição de termos que apresentam mineração em grandes massas de dados. A chave é composta da seguinte forma:

(Swarm Intelligence OR Evolutionary Computation OR Bio Inspired Algorithm) AND (Hadoop MapReduce OR Hadoop Map Reduce) AND (Data OR Mining)

Dada a relevância na área da Ciência da Computação, os repositórios de busca utilizados foram: ACM Digital Library, IEEE Xplore, ScienceDirect e Scopus. Após a busca com a chave nos repositórios citados foi aplicado os critérios de seleção de artigos citados a seguir.

A Tabela 1 apresenta a quantidade de artigos retornados na pesquisa com a chave de busca em cada repositório, totalizando 237 artigos.

Tabela 1. Resultado da busca com a chave de pesquisa em cada repositório

IEEE	*ACM*	*ScienceDirect*	*Scopus*
168	52	10	7

A Tabela 2 exibe os critérios de inclusão e exclusão dos artigos encontrados na busca. Se o artigo caracterizar a presença de pelo menos um critério de exclusão, esse artigo não será utilizado na revisão.

Tabela 2. Critérios de Inclusão e Exclusão

Critérios de Inclusão	**Critérios de Exclusão**
CI 1 - Artigo propõe alguma aplicação de software que realiza mineração em grandes volumes de dados. Além disso, o mesmo deve envolver algum algoritmo bio-inspirado projetado em Hadoop MapReduce.	CE1 - Artigos duplicados. CE2 - Artigos que não estejam escritos em inglês. CE3 - Artigos publicados há mais de 10 anos.

No intuito de garantir a qualidade dessa revisão é necessário analisar cada artigo selecionado considerando os critérios citados. Assim, em cada artigo selecionado será necessário analisar as seguintes partes: Título, Resumo, Palavras chave, Mecanismo proposto, Resultados e Conclusão.

Após aplicar os critérios de inclusão e exclusão nos artigos encontrados com a chave de busca foram obtidos 8 artigos, conforme Tabela 3.

Com o intuito de complementar a busca, a pesquisa por trabalhos relacionados foi realizada diretamente em alguns *Journals* importantes para a área. Dessa forma, foi encontrado e incluído nesta revisão um artigo científico de cada um dos seguintes *Journals*: *Scientific Research and Essays*, *International Journal on Recent and Innovation Trends in Computing and*

Tabela 3. Trabalhos Selecionados por repositório

IEEE	*ACM*	*ScienceDirect*	*Scopus*
4	2	0	2

Communication, *International Journal of Scientific & Technology Research Volume* e *Mathematical Problems in Engineering*. Todos os 4 trabalhos encontrados possuem os critérios de inclusão e não possuem critérios de exclusão. Desta maneira, no total, foram selecionados 12 artigos científicos para compor esta revisão.

Para cada artigo selecionado, considerando todo o processo de seleção, este deverá ser analisado conforme procedimento de avaliação, de tal forma que seja possível extrair do artigo as seguintes informações: Tarefa ou etapa de mineração de dados; Algoritmo bio-inspirado utilizado; Se a base utilizada está disponível; E quais V's estão sendo considerados.

6 Trabalhos Revisados

Esta seção apresenta os trabalhos encontrados na literatura que empregam algoritmos bio-inspirados projetados no *framework Hadoop MapReduce* aplicados em mineração de grandes massas de dados. Após todo processo de seleção já descrito, foram selecionados 12 artigos diretamente relacionados com o escopo deste trabalho. O objetivo dessa revisão é identificar as principais contribuições que esses trabalhos realizaram em mineração de grandes massas de dados, tornando possível obter uma visão mais apurada dos recursos e dificuldades existentes atualmente nessa área. Como todos os trabalhos utilizam o *framework Hadoop*, a paralelização com tolerância a falhas e balanceamento de carga é uma característica implícita e comum em todos os trabalhos comentados a a seguir.

Aljarah (2012) [5] propôs um algoritmo que realiza a tarefa de agrupamento e tem como base o algoritmo *PSO* adaptado ao paradigma *MapReduce*. O algoritmo é composto por três módulos principais que são executados sequencialmente a cada iteração. No primeiro módulo ocorre a atualização do centróide de cada partícula do enxame através de um *job MapReduce* no arquivo que mantem o enxame. No segundo módulo ocorre outro *job MapReduce* nos grandes arquivos de dados e faz o cálculo do *fitness* considerando os novos centróides. No terceiro módulo ocorre a combinação das informações atualizadas no primeiro e segundo módulos, além de calcular o melhor resultado individual de cada partícula e o melhor resultado global do enxame finalizando a iteração do algoritmo. As bases *MAGIC Gamma Telescope Data Set* (19.020 instâncias de dados, 10 dimensões, atributo-meta com 2 classes e 3 *Megabytes* de tamanho), *Poker Hand Data Set* (1.025.010 instâncias de dados, 10 dimensões, atributo-meta com 10 classes e 49 *Megabytes* de tamanho) e *Covertype Data Set* (581.012 instâncias de dados, 54 dimensões, atributo-meta com 7 classes e 199.2 *Megabytes*

de tamanho) disponíveis no repositório *UCI*[4] foram utilizadas nesse trabalho. Também foi utilizada a base *Electricity* (45.312 instâncias de dados, 8 dimensões, atributo-meta com 2 classes e 6 *Megabytes* de tamanho) que está disponível no *MOA*[5]. Os resultados mostraram uma melhor qualidade comparando com o algoritmo *K-Means* além de possuir uma boa relação de desempenho (*speedup*) mantendo a qualidade do agrupamento. Esse mesmo algoritmo foi utilizado em Aljarah (2013) [6] como um componente que realiza agrupamento em volume de dados de tráfico de redes de grande escala. O algoritmo compõe uma arquitetura de sistema de detecção de intrusos. Os dados analisados são coletados de tráfego de redes. Com essa arquitetura os resultados obtiveram uma boa taxa de detecção de invasão verdadeira e baixa taxa de detecção falsa. Além disso, foi mantida uma boa escalabilidade e *speedup*. O mecanismo proposto nos dois trabalhos distribui o cálculo da velocidade e dos centróides de cada partícula no primeiro *job MapReduce*. O segundo *job* calcula a média das distâncias e, por último, a função *fitness* é executada em uma rotina sequencial complementar.

Judith (2015) [27] apresenta um sistema que realiza tarefa de agrupamento em documentos utilizando o algoritmo *PSO*. Além de *PSO*, o sistema proposto também utiliza o algoritmo de agrupamento de dados *K-Means* e a técnica de redução de dimensionalidade *Latent Semantic Indexing (LSI)*. A utilização do algoritmo *PSO* projetado no paradigma *MapReduce* visa encontrar melhores valores para os centróides explorando assim a característica vantajosa da busca global. Na fase de mapeamento o *fitness* é calculado considerando a distância dos documentos até o centróide do *cluster*. Na fase de redução os valores dos centróides são atualizados. Após isso, o *K-Means* projetado em *MapReduce* é executado visando melhorar a qualidade dos *clusters*. Na sequência, a técnica *LSI* é aplicada como método de redução de dimensionalidade dos documentos melhorando a eficiência do processo de agrupamento. A base de documentos *Reuters-21578* (21578 documentos, 131 classes, 42686 termos e tamanho de 28 *Megabytes*) foi utilizada nos experimentos e está disponível no *UCI* (http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html). Também foi utilizada a base de dados RV1 (1000000 documentos, 860 classes, 3268650 termos). A qualidade do agrupamento de dados do sistema proposto foi melhor que a qualidade obtida pelo *K-Means* executado isoladamente. *K-Means* é fortemente dependente dos centróides iniciais e converge facilmente ao ótimo local. Assim, em vez de inicializar os centróides com valores aleatórios é utilizado o PSO. Dessa forma a qualidade dos *clusters* melhora. A utilização da técnica *LSI* também impactou na melhora da qualidade dos *clusters*. O sistema também demonstrou bom *speedup* com a utilização do *framework Hadoop MapReduce*.

Al-Mad (2014) [3] apresenta um algoritmo que realiza a tarefa de agrupamento e tem como base o algoritmo GSO. A cada iteração do algoritmo GSO ocorre a execução de um

[4] *UCI Machine Learning Repository* é um repositório de dados utilizados pela comunidade de aprendizado de máquinas para análise empírica de algoritmos, https://archive.ics.uci.edu/ml/index.html.
[5] *Massive Online Analysis (MOA)* é um *framework* de mineração de dados que disponibiliza algumas bases de dados para realizar analises de mineração de dados. http://moa.cms.waikato.ac.nz/datasets/.

job MapReduce no arquivo de dados. Nesse processo, cada indivíduo procura apenas um centróide como subsolução no mapeamento. Na redução são combinadas as subsoluções compondo a solução global. As bases utilizadas foram: *MAGIC Gamma Telescope Data Set*, *Poker Hand Data Set* e *Covertype Data Set* disponíveis no *UCI*. Também foi utilizada a base *Electricity* que está disponível no *MOA*. Os resultados mostram que esse algoritmo possui um bom tempo de performance e mantem melhor qualidade nos grupos formados do que o *K-Means* e pelo algoritmo proposto por Aljarah (2012) [5]. Nesse mecanismo, o mapeamento calcula a distância de cada indivíduo do enxame GSO para cada objeto de dados distribuído no paradigma *MapReduce*. Após finalizar a execução do *job MapReduce* é atualizado o *fitness* e a luciferina.

O algoritmo *ACO* foi proposto por Bhavani (2014) [8] para realizar a tarefa de agrupamento. Para a aplicação do ACO, em uma etapa de pré-processamento, antes da execução do *job MapReduce*, um grafo é construído a partir dos dados que serão agrupados. No processo de mapeamento a colônia de formigas encontra a árvore de cobertura mínima do grafo e na redução ocorre a atualização do feromônio tendo como base a média da menor árvore de cobertura mínima. Ao final de todas as iterações ocorre a poda da árvore que utiliza a quantidade de feromônio encontrando no final do processo. Quanto menos feromônio maior é a probabilidade da poda. Em relação aos dados analisados, foi citado apenas o tamanho de três bases: 15,7 *Kilobytes*, 4,2 *Megabytes* e 17,8 *Megabytes*. Esse algoritmo conseguiu alcançar um bom *speedup*. O mecanismo distribui o algoritmo ACO de tal forma que no mapeamento a quantidade de feromônio influencia no cálculo da arvore de cobertura mínima. Já na redução o feromônio é atualizado.

Outro trabalho que utiliza *ACO* para realizar a tarefa de agrupamento foi apresentado por Yang (2012) [55]. No sistema apresentado, antes de executar o *job MapReduce* o componente global inicializa os parâmetros, gerencia os movimentos e posições dos agentes utilizando coordenadas. Dessa forma, a cada iteração após o movimento das formigas realizando o agrupamento de dados é executado o *job MapReduce*. No mapeamento ocorre o calculo de similaridade dos objetos de dados. Na redução ocorre o média global de similaridade. Nos experimentos foram utilizadas as bases Iris (150 instâncias de dados, 4 dimensões, atributo-meta com 3 classes, 4,4 *Kilobytes* de tamanho), Ecoli (336 instâncias de dados, 8 dimensões, atributo-meta com 8 classes, 19 *Kilobytes* de tamanho) e *Letter Recogination* (20,000 instances, 26 classes, 17 attributes) disponíveis no *UCI*. Foi comprovado que o tempo de execução do sistema em bases grandes diminui à medida que é incrementado o número de nodos que compõe o *cluster*. Entretanto, percebe-se que o tempo de execução do sistema em pequenas bases de dados aumenta a medida que é incrementado o número de nodos que compõe o *cluster* devido ao tempo consumido na comunicação entre os nodos na rede. Em relação a qualidade do agrupamento, foi comparado os resultados da versão do sistema com e sem a arquitetura *MapReduce* e não foi possível concluir melhoria na qualidade dos grupos.

Hans (2015) [25] apresenta um aplicativo de *software* que realiza a tarefa de agrupamento de dados utilizando *GA*. Na fase de mapeamento é calculado o *fitness* dos indivíduos considerando utilizando a inter e intra distancia dos *clusters*. Na fase de redução, novos indivíduos com novos centróides gerados e analisados. Dessa forma, a cada iteração o *job MapReduce* é executado. A base *Europe* (169308 instâncias de dados, 2 dimensões, atributo-meta com 2 classes, 3,6 *Megabytes* de tamanho) disponíveis no *SIPU*[6] foi utilizada nesse trabalho. Foi comparado a versão do aplicativo executado de forma sequencial com a versão projetada em *Hadoop MapReduce*. Foi demonstrado que os resultados de qualidade dos *clusters* foram semelhantes. Em relação ao tempo de execução, a versão do aplicativo projetado em *Hadoop MapReduce* consumiu menos tempo.

O algoritmo *GA* também foi utilizado para realizar a etapa de seleção de atributos antes de realizar a geração de regras de classificação em Ferrucci (2015) [20]. Como o problema de seleção de atributos otimiza os dados de treinamento para realizar a tarefa de classificação esse trabalho foi considerado como tarefa de classificação. Para tal, foi construído um *framework* de desenvolvimento que paraleliza o algoritmo *GA* utilizando *Hadoop MapReduce*. O *framework* adapta o paradigma *MapReduce* de tal forma que é possível executar em uma única iteração vários processos de mapeamento antes e depois da redução. Dessa forma foi possível mapear todos os processos que ocorrem em uma geração de indivíduos em *GA* em um único *job MapReduce*. O processamento da população de indivíduos é realizada de forma segmentada no modelo de ilhas e o GA é executado de forma independente em cada ilha. Dessa forma o algoritmo mantém várias populações diferentes sendo assim multipopulacional. Além disso, ocorre migração de indivíduos entre as ilhas de acordo com a parametrização, favorecendo a co-evolução. As bases utilizadas disponíveis no *UCI* são: *Statlog (German Credit Data) Data Set* (1000 instâncias de dados, 20 dimensões, atributo-meta com 2 classes, 78 *Kilobytes* de tamanho) e *Audiology Data Set* (226 instâncias de dados, 69 dimensões, atributo-meta com 24 classes, 32 *Kilobytes* de tamanho). Os resultados demonstram que a precisão e tempo de execução foram melhores para a aplicação desenvolvida como base no *framework Hadoop MapReduce* em comparação com o *Weka*[7] e com uma versão pseudo-distribuída com apenas uma ilha.

Em Peralta (2015) [45], também é abordada a etapa de seleção de atributos da tarefa de classificação de dados. Nesse caso, o algoritmo bio-inspirado utilizado é uma variação do *GA*. Diferente de *GA*, esse algoritmo verifica a similaridade entre dois possíveis indivíduos antes de aplicar o operador *crossover*. O sistema proposto é composto por duas fases principais. Cada fase representa um *job MapReduce*. Na primeira fase, o algoritmo é executado na fun-

[6] *Speech and Image Processing Unit* é um grupo de pesquisa da *School of Computing University of Eastern Finland* que trabalha com padrões recognitivos, compressão e mineração de dados. Dentre outros benefícios à comunidade científica, algumas bases de dados para realizar a tarefa de agrupamento são disponibilizadas, http://cs.joensuu.fi/sipu/datasets/.

[7] *Weka* é uma coleção de algoritmos de aprendizado de máquina utilizados em tarefas de mineração de dados. Os algoritmos aplicados diretamente em um conjunto de dados (*dataset*) ou ser chamado de um código *Java* [20].

ção de mapeamento produzindo como saída um vetor binário. Esse vetor possui informações de relevância de cada atributo dos dados mapeados. Na função de redução, calcula-se a média dos vetores emitidos na saída do mapeamento e com isso é obtido o vetor final, chamado de vetor *x*. O vetor *x* define quais atributos serão mantidos ou não na base de treinamento para realização da tarefa de classificação. No segundo *job MapReduce*, o vetor *x* é utilizado para remover os atributos menos relevantes na base de dados. Na fase de mapeamento é verificado para cada atributo de dados se o mesmo é relevante ou não de acordo com o vetor *x*. Se o atributo não for relevante o mesmo não é emitido para a fase de redução, assim esse atributo é removido da base de treinamento. Caso contrário, o atributo é mantido na base de dados de treinamento. Na função de redução os dados são apenas concatenados. Após a execução do segundo *job MapReduce*, a base de treinamento possuirá apenas dados com os atributos relevantes. Com isso, a base de treinamento reduzida é utilizada por algoritmos de classificação. Foram utilizadas duas bases nos experimentos. Uma base (500000 instâncias de dados, 2000 dimensões, o autor não explicitou mais informações) foi gerada artificialmente pelo *Pascal Large Scale Learning Challenge* (http://largescale.ml.tu-berlin.de/instructions/), essa base não está disponível. A outra base de dados (32000000 instâncias de dados, 631 dimensões, atributo-meta com 2 classes, 56 *Gigabytes* de tamanho) está disponível na página *web* do evento *Evolutionary Computation for Big Data and Big Learning Workshop* realizado em 2014 (http://http://cruncher.ncl.ac.uk/bdcomp/). Foi comparado o tempo de execução do sistema projetado em *Hadoop MapReduce* com sua versão sequencial. O resultados demonstram que o tempo de execução da versão sequencial cresce numa tendência quase exponencial a medida que o tamanho da base de dados aumenta. Foi percebido também que a precisão dos algoritmos tradicionais utilizados na classificação obtiveram melhor precisão na tarefa de classificação utilizando a base de dados após a etapa de seleção de atributos.

Al-Madi (2013) [4] apresenta um algoritmo que realiza classificação e tem como base a implementação o algoritmo *GP*. No mapeamento é calculado o *fitness* de cada programa. Na redução ocorre a seleção, *crossover* e mutação, além de gerar nova população. Foram utilizados seis *Datasets* disponíveis na UCI, porém, o artigo não deixa claro quais foram as bases. Resultados mostram que foi obtida uma alta precisão de classificação além de manter escalabilidade e diminuir o tempo de execução por mapeamento.

Daoudi (2014) [14] apresenta uma estratégia para realizar a tarefa de agrupamento. Esse mecanismo utiliza o algoritmo *DE* e é composto por três níveis. No primeiro nível ocorre processamento *MapReduce* onde atuam a mutação e seleção que calculam novos centróides. No segundo nível ocorre outro processo *MapReduce* que calcula a menor distância entre o registro e os centróides avaliando assim os centróides gerados na fase anterior. Por último, ocorre um processo de atualização do *fitness* e seleção dos indivíduos. Foram utilizadas as bases: armstrong-v1 (12.582 instâncias de dados, atributo-meta com 2 classes, 274 *Kilobytes* de tamanho), armstrong-v2 (12.582 instâncias de dados, atributo-meta com 3 classes, 274 *Kilobytes* de tamanho), bhattacharjee (12.600 instâncias de dados, atributo-meta com 5 clas-

ses, 1,7 *Megabytes* de tamanho), chowdary (22.283 instâncias de dados, atributo-meta com 2 classes, 88 *Kilobytes* de tamanho), dyrskjot (7.129 instâncias de dados, atributo-meta com 3 classes, 242 *Kilobytes* de tamanho), golubv1 (7.129 instâncias de dados, atributo-meta com 2 classes, 270 *Kilobytes* de tamanho), golubv2 (7.129 instâncias de dados, atributo-meta com 3 classes, 270 *Kilobytes* de tamanho), laiho (22.883 instâncias de dados, atributo-meta com 2 classes, 406 *Kilobytes* de tamanho), nutt-v1 (12.625 instâncias de dados, atributo-meta com 4 classes, 403 *Kilobytes* de tamanho), nutt-v2 (12.625 instâncias de dados, atributo-meta com 2 classes, 403 *Kilobytes* de tamanho), nutt-v3 (12.625 instâncias de dados, atributo-meta com 2 classes, 403 *Kilobytes* de tamanho) , pomeroy-v1 (7.129 instâncias de dados, atributo-meta com 2 classes, 108 *Kilobytes* de tamanho), pomeroy-v2 (7.129 instâncias de dados, atributo-meta com 5 classes, 108 *Kilobytes* de tamanho), ramaswamy (16.063 instâncias de dados, atributo-meta com 14 classes, 835 *Kilobytes* de tamanho), shipp-2002-v1 (7.129 instâncias de dados, atributo-meta com 3 classes, 217 *Kilobytes* de tamanho), su (12.533 instâncias de dados, atributo-meta com 10 classes, 939 *Kilobytes* de tamanho), west (7.129 instâncias de dados, atributo-meta com 2 classes, 212 *Kilobytes* de tamanho) e Yeoh-v2 (12.625 instâncias de dados, atributo-meta com 6 classes, 2,5 *Megabytes* de tamanho). As mesmas estão disponíveis em *Schliep lab*[8] . Os resultados comprovam que esse modelo é mais preciso que *K-Means* e é mais estável e robusto que Aljarah (2012) [5] pois possui melhor desempenho a medida que o volume da base de dados aumenta.

Bhavani (2011) [9] apresenta duas estratégias para realizar agrupamento. Uma envolve *DE* e *K-Means* e a outra envolve o uso de *DE*, *ACO* e *K-Means*. Em ambas estratégias o algoritmo *DE* é utilizado para realizar a atualização global dos centróides enquanto que *K-Means* é utilizado localmente. Entretanto, no primeiro mecanismo o limiar de ativação dos centróides é atualizado de forma randômica e é verificada uma deterioração da velocidade de convergência dos grupos. Esse problema é contornado no segundo mecanismo com a utilização do *ACO* para realizar a atualização do limiar. Em relação as bases utilizadas, esse artigo menciona apenas que as bases são da área da bioinformática. Assim, o segundo mecanismo obteve resultados com melhor qualidade e desempenho. Ambos mecanismos aplicam o algoritmo DE apenas na redução, mas no segundo mecanismo o algoritmo o *ACO* é distribuído de tal forma que o feromônio seja analisado no mapeamento e é atualizado na redução.

Na próxima Seção é apresentada a discussão comparativa entre os trabalhos descritos considerando alguns aspectos importantes.

[8] *Schliep lab bioinformatics*, é o laboratório do Departamento de Bioinformática da Freie Universität Berlin que disponibiliza alguns *Datasets* dessa área para serem utilizados por algoritmos de aprendizado de máquina. http://bioinformatics.rutgers.edu/.

7 Discussão

Nessa seção, serão discutidos alguns critérios que visam identificar as semelhanças e diferenças entre os 12 trabalhos apresentados na seção anterior. Quatro critérios foram analisados. São eles: tarefa de mineração de dados abordada, tipo de algoritmo bio-inspirado utilizado, como o algoritmo bio-inspirado foi projetado no *Hadoop MapReduce*, a disponibilidade das bases utilizadas e quais V's foram abordados. A distribuição dos dados minerados em *HDFS* é uma característica comum a todos os trabalhos.

7.1 Tarefa de mineração de dados

Esse critério busca evidenciar qual é a proporção de cada tarefa de mineração de dados abordada considerando todos os trabalhos levantados na revisão. Como visto na Figura 2, pesquisas que realizam agrupamento de dados são mais expressivas que os trabalhos que realizam a tarefa de classificação.

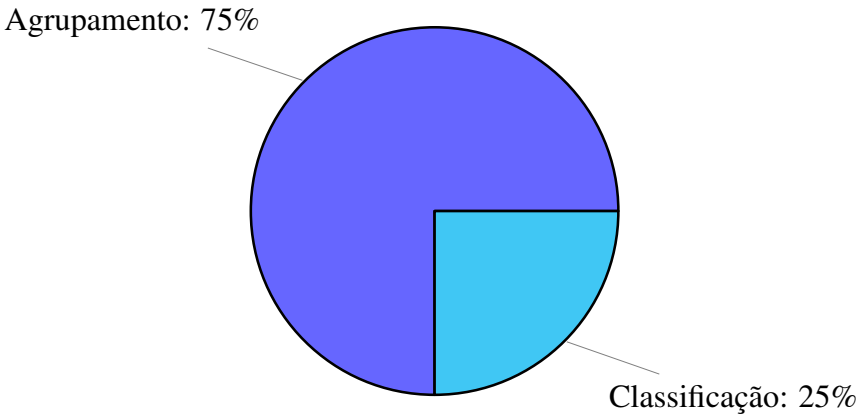


Figura 2. Distribuição de tarefas de mineração

A proporção de trabalhos que realizam agrupamento é bem mais significante do que a proporção que aborda a tarefa de classificação. Isso pode ser atribuído ao fato de que abordar a tarefa de agrupamento utilizando o paradigma *MapReduce* se torna um processo onde as características do problema se beneficiam das características do paradigma, e vice-versa. A computação necessária para realizar agrupamento é, basicamente, o cálculo da distância entre o centróide representado por uma possível solução (indivíduo ou partícula, por exemplo) e o dado que está sendo agrupado. Mesmo que todo o arquivo de dados seja mapeado de forma independente e distribuída, o cálculo da distância possui todas as informações necessárias para ser efetuado. Essa tarefa se adequa bem ao processo de mapeamento pois todo volume de dados será processado de forma independente a cada iteração. Já na tarefa de classifica-

ção, é necessário armazenar e controlar a frequência de ocorrência dos dados. Isso deixa o mecanismo mais complexo devido ao controle que será necessário para analisar a frequência de dados, visto que o processo de mapeamento analisa cada objeto do volume de dados de forma independente.

Percebe-se que apenas duas tarefas de mineração de dados foram abordadas nos trabalhos revisados. Considerando o protocolo de pesquisa e repositório de artigos científicos utilizados nessa pesquisa, não foi encontrado nenhum trabalho que aborde tarefas como Associação e Regressão. Portanto, essas tarefas possuem vasto campo de exploração considerando a linha de solução abordada nessa revisão.

7.2 Tipos de algoritmos bio-inspirados

Este critério visa sumarizar a variedade de algoritmos bio-inspirados encontrados dentro do contexto da revisão. A grande diversidade de algoritmos bio-inspirados aplicados em mineração de dados utilizando *Hadoop MapReduce* pode ser evidenciada na Figura 3.

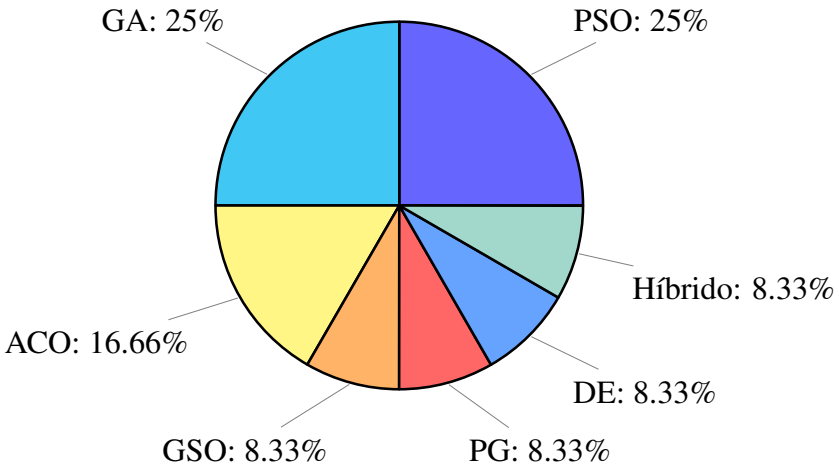


Figura 3. Distribuição de tipos de algoritmos bio-inspirados

Observa-se que *PSO* e *GA* são mais utilizados por serem os algoritmos mais conhecidos pela busca eficiente de solução global. Além de terem bom desempenho em problemas de domínio contínuo, sua fácil compreensão e desenvolvimento são características interessantes que favorecem suas escolhas.

De maneira geral, cada pesquisa utiliza as vantagens particulares de cada algoritmo bio-inspirado, explorando a diferenciada capacidade de diversificação e intensificação na exploração do espaço de solução. Apenas um trabalho utiliza uma abordagem híbrida com mais de um algoritmo bio-inspirado, *DE* e *ACO*. Existem outros algoritmos bio-inspirados como

Colônia de Abelhas Artificiais (*Artificial Bee Colony - ABC*), Algoritmo do Morcego (*Bat Algorithm - BA*), Algoritmo de Vagalumes (*Firefly Algorithm - FA*) que possuem potencial de serem utilizados com *Hadoop MapReduce* em mineração de *Big Data*.

7.3 Projeto do *Hadoop MapReduce*

Três modelos de paralelismo foram identificados e são descritos a seguir.

A principal característica do Modelo 1 é a execução de apenas um *job MapReduce* por ciclo de iteração do algoritmo (conforme Figura 4, onde o ícone com o desenho do planeta Terra representa o algoritmo bio-inspirado). Além disso, cada execução do *job MapReduce* processa toda a população de indivíduos no mapeamento do volume de dados. Como apresentado na Figura 4, o algoritmo bio-inspirado está projetado no *job Hadoop MapReduce*, portanto será paralelizado. Sete trabalhos usam esse modelo.

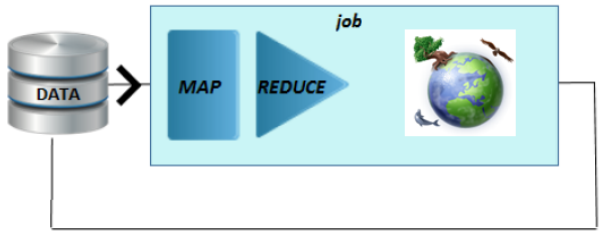


Figura 4. Modelo 1:Um *job MapReduce*

O Modelo 2 tem como principal característica a execução de dois *jobs MapReduce* executados um após o outro (Rótulos 1 e 2 da Figura 5). Logo em seguida, é executada uma rotina que finaliza o ciclo de iteração do algoritmo (Rótulo 3 da Figura 5).

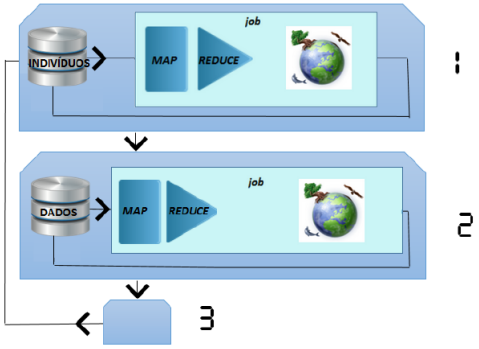


Figura 5. Modelo 2: Dois *jobs MapReduce*

Conforme Figura 6, o primeiro *job MapReduce* processa dados que representam a

movimentação dos indivíduos da população no espaço de solução. Já o segundo *job MapReduce* processa o volume de dados que está sendo minerado considerando a população de indivíduos. Na sequência uma rotina é executada para realizar as últimas operações de um ciclo do algoritmo bio-inspirado. Esse modelo processa a população de indivíduos de forma paralela através do *job MapReduce* representado pelo Rótulo 1. Quatro trabalhos utilizam esse modelo. Apesar de [27] não realizar a execução de um *job MapReduce* no arquivo com a população de indivíduos, o mesmo foi classificado no Modelo 2 por executar dois *jobs MapReduce*, um após o outro, que corresponde a principal característica desse modelo.

O Modelo 3 tem como principal característica a execução de um *job MapReduce* onde múltiplas funções de mapeamento são processadas antes e depois da função de redução (Figura 6).

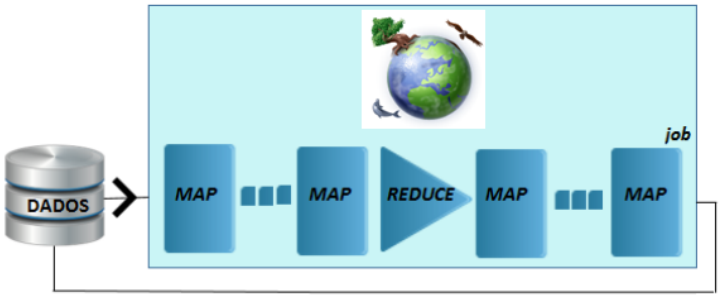


Figura 6. Modelo 3: *Job com múltiplos Maps e Reduce*

Esse modelo consegue executar em um único *job MapReduce* em várias fases do algoritmo bio-inspirado. Cada fase do algoritmo pode ser executada na função de mapeamento ou de redução. A desvantagem é a complexidade necessária para alterar seu *framework Hadoop MapReduce* e assim alterar o comportamento padrão do *job MapReduce* que é executar apenas um mapeamento e uma redução. Apenas um trabalho utilizou esse modelo.

A Figura 7 exibe a sumarização dos trabalhos revisados agrupados pelo modelo que cada trabalho melhor se assemelha.

O Modelo 1 tende a ser mais vantajoso por ser mais simples de ser desenvolvido e por conseguir utilizar os benefícios do *framework MapReduce* no processamento dos dados que estão sendo minerados. Esse modelo é o mais utilizado nos trabalhos revisados. Como não é vantajoso executar o *job MapReduce* em arquivos pequenos [55] [28], como ocorre no arquivo que armazena a população de indivíduos no Modelo 2, este modelo torna o desenvolvimento mais trabalhoso sem ganho de desempenho com a paralelização devido ao tempo consumido na comunicação entre os nodos do *cluster*. Dessa forma, a ocorrência do Modelo 2 nos trabalhos revisados foi menor que a ocorrência do Modelo 1. Apenas um trabalho foi classificado no Modelo 3 devido a particularidade não muito comum de envolver alteração

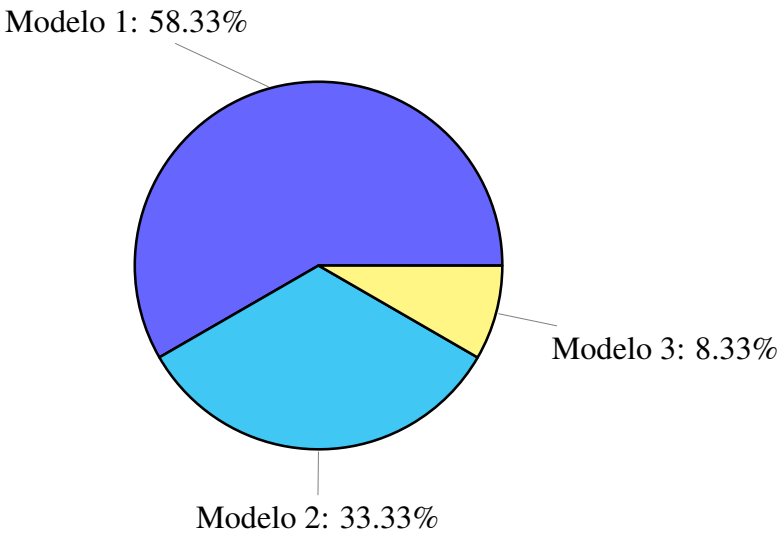


Figura 7. Modelos de projeto *Hadoop MapReduce*

no comportamento padrão do *framework Hadoop MapReduce*.

7.4 Disponibilidade de Base de Dados e V's abordados

A grande maioria dos trabalhos revisados, 75%, utilizam bases disponíveis em repositórios na *web*. Os repositórios mais utilizados são: *UCI Machine Learning Repository e Massive Online Analysis - MOA*.

A disponibilidade dessas bases é muito importante para que seja possível comparar diferentes mecanismos e assim conseguir avaliar as vantagens e desvantagens entre eles.

Em relação aos 5 V's que caracterizam *Big Data*, o único V abordado nos trabalhos é o Volume. Ainda assim, o Volume de dados minerado nos trabalhos não chegam a medir em *Petabytes*, que corresponde a 10^{15} *Bytes*, conforme a definição dessa característica. As bases de dados utilizadas estão em *Megabytes*, que corresponde a 10^{6} *Bytes*. Apenas uma base utilizada alcançou a grandeza de *Gigabytes*. O conceito de *Big Data* menciona dados com volume em *Terabytes*. Apesar disso, percebe-se que os trabalhos encontrados na literatura ainda não conseguiram colocar em prática a manipulação de dados em *Terabytes*.

8 Conclusão

Esse artigo apresentou a revisão de trabalhos que realizam mineração em grandes massas de dados utilizando algoritmos bio-inspirados embasados no *framework Hadoop MapReduce*. Essa combinação pode proporcionar soluções de mineração de dados mais adequadas ao contexto de grandes massas de dados. Com isso, consegue-se entender e perceber o quão flexível é projetar algoritmos dessa natureza no *framework Hadoop MapReduce*. Isso porque é simples e flexível adaptar a execução do algoritmo bio-inspirado na função *map*, *reduce* ou em ambas. Dessa forma obtém-se paralelização, tolerância a falhas e escalabilidade embasada no *framework MapReduce* com o grande poder de otimização fornecido pelos algoritmos bio-inspirados.

Entretanto, pode-se perceber que a execução do *job MapReduce* em arquivos de dados relativamente pequenos não é recomendado devido ao tempo consumido na comunicação entre os nodos do *cluster*.

Na busca eficiente por resultados de mineração de dados, características como a intensificação e a diversificação no espaço de busca pertencentes aos algoritmo bio-inspirados, são responsáveis por proporcionar às pesquisas desenvolvidas a encontrar as melhores soluções no espaço de busca.

Também percebeu-se a disponibilidade de bases na *web* que foram utilizadas nos experimentos dos trabalhos revisados. Isso favorece o processo de avaliação comparativa entre os trabalhos existentes e identifica *benchmarks* para desenvolvimentos futuros.

Com relação às características definidas pela contextualização de *Big Data*, nem todos os V's são abordados pelos trabalhos encontrados. Todos trabalhos abordam apenas o volume dentre as 5 propriedades *Big Data*. Espera-se que, com a evolução contínua do estado da arte, progressivamente outras características *Big Data* sejam abordadas na proposta de solução dos trabalhos futuros de tal forma que todos os 5 V's sejam abordados.

Contribuição dos autores:

- Sandro Roberto Loiola de Menezes: Pesquisa e escrita do manuscrito.

- Rebeca Schroeder Freitas: Correções e direcionamentos da pesquisa.

- Rafael Stubs Parpinelli: Coordenação do trabalho, correções e direcionamentos da pesquisa.

Referências

[1] Big data defintion, gartner, inc. `http://www.gartner.com/it-glossary/big-data/`. Accessed: 2015-08-07.

[2] R. Al-Khannak and B. Bitzer. Grid computing as an innovative solution for power system's reliability and redundancy. In *Clean Electrical Power, 2009 International Conference on*, pages 790–797. IEEE, 2009.

[3] N. Al-Madi, I. Aljarah, and S. A. Ludwig. Parallel glowworm swarm optimization clustering algorithm based on mapreduce. In *Swarm Intelligence (SIS), 2014 IEEE Symposium on*, pages 1–8. IEEE, 2014.

[4] N. Al-Madi and S. A. Ludwig. Scaling genetic programming for data classification using mapreduce methodology. In *Nature and Biologically Inspired Computing (NaBIC), 2013 World Congress on*, pages 132–139. IEEE, 2013.

[5] I. Aljarah and S. A. Ludwig. Parallel particle swarm optimization clustering algorithm based on mapreduce methodology. In *Nature and Biologically Inspired Computing (NaBIC), 2012 Fourth World Congress on*, pages 104–111. IEEE, 2012.

[6] I. Aljarah and S. A. Ludwig. Mapreduce intrusion detection system based on a particle swarm optimization clustering algorithm. In *Evolutionary Computation (CEC), 2013 IEEE Congress on*, pages 955–962. IEEE, 2013.

[7] A. Asbern and P. Asha. Performance evaluation of association mining in hadoop single node cluster with big data. In *Circuit, Power and Computing Technologies (ICCPCT), 2015 International Conference on*, pages 1–5. IEEE, 2015.

[8] R. Bhavani and G. S. Sadasivam. A novel ant based clustering of gene expression data using mapreduce framework. *International Journal on Recent and Innovation Trends in Computing and Communication*, 2:398–402, 2014.

[9] R. Bhavani, G. S. Sadasivam, and R. Kumaran. A novel parallel hybrid k-means-deaco clustering approach for genomic clustering using mapreduce. In *Information and Communication Technologies (WICT), 2011 World Congress on*, pages 132–137. IEEE, 2011.

[10] S. Binitha and S. S. Sathya. A survey of bio inspired optimization algorithms. *International Journal of Soft Computing and Engineering*, 2(2):137–151, 2012.

[11] M. Chen, S. Mao, and Y. Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.

[12] S. Cheng, Y. Shi, Q. Qin, and R. Bai. Swarm intelligence in big data analytics. In *Intelligent Data Engineering and Automated Learning–IDEAL 2013*, pages 417–426. Springer, 2013.

[13] P.-T. Chung and S. H. Chung. On data integration and data mining for developing business intelligence. In *Systems, Applications and Technology Conference (LISAT), 2013 IEEE Long Island*, pages 1–6. IEEE, 2013.

[14] M. Daoudi, S. Hamena, Z. Benmounah, and M. Batouche. Parallel diffrential evolution clustering algorithm based on mapreduce. In *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, pages 337–341. IEEE, 2014.

[15] L. Davis et al. *Handbook of genetic algorithms*, volume 115. Van Nostrand Reinhold New York, 1991.

[16] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[17] Y. Demchenko, C. de Laat, and P. Membrey. Defining architecture components of the big data ecosystem. In *Collaboration Technologies and Systems (CTS), 2014 International Conference on*, pages 104–112. IEEE, 2014.

[18] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey. Addressing big data issues in scientific data infrastructure. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, pages 48–55. IEEE, 2013.

[19] K. A. Doshi, T. Zhong, Z. Lu, X. Tang, T. Lou, and G. Deng. Blending sql and newsql approaches: reference architectures for enterprise big data challenges. In *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference on*, pages 163–170. IEEE, 2013.

[20] F. Ferrucci, P. Salza, M. T. Kechadi, and F. Sarro. A parallel genetic algorithms framework based on hadoop mapreduce. pages 1664–1667, 2015.

[21] I. Fister Jr, X.-S. Yang, I. Fister, J. Brest, and D. Fister. A brief review of nature-inspired algorithms for optimization. *Electrotechnical Review*, 80(3), 2013.

[22] A. Ghosh and B. Nath. Multi-objective rule mining using genetic algorithms. *Information Sciences*, 163(1):123–133, 2004.

[23] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[24] J. Han, M. Kamber, and J. Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

[25] N. Hans, S. Mahajan, and S. Omkar. Big data clustering using genetic algorithm on hadoop mapreduce. *International Journal of Scientific Technology Research*, 4, 2015.

[26] C. Jin, C. Vecchiola, and R. Buyya. Mrpga: an extension of mapreduce for parallelizing genetic algorithms. In *eScience, 2008. eScience'08. IEEE Fourth International Conference on*, pages 214–221. IEEE, 2008.

[27] J. Judith and J. Jayakumari. An efficient hybrid distributed document clustering algorithm. *Scientific Research and Essays*, 10(1):14–22, 2015.

[28] A. Katal, M. Wazid, and R. Goudar. Big data: Issues, challenges, tools and good practices. In *Contemporary Computing (IC3), 2013 Sixth International Conference on*, pages 404–409. IEEE, 2013.

[29] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, 51(1):7–15, 2009.

[30] P. Koturwar, S. Girase, and D. Mukhopadhyay. A survey of classification techniques in the area of big data. *International Journal of Advance Foundation and Research in Computer*, 1, 2015.

[31] J. R. Koza. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press, 1992.

[32] K. Krishnanand and D. Ghose. Detection of multiple source locations using a glowworm metaphor with applications to collective robotics. In *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE*, pages 84–91. IEEE, 2005.

[33] M. Kutlu, G. Agrawal, and O. Kurt. Fault tolerant parallel data-intensive algorithms. In *High Performance Computing (HiPC), 2012 19th International Conference on*, pages 1–10. IEEE, 2012.

[34] Y. Lai and S. ZhongZhi. An efficient data mining framework on hadoop using java persistence api. In *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, pages 203–209. IEEE, 2010.

[35] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of massive datasets*. Cambridge University Press, 2014.

[36] H. Liu and H. Motoda. *Feature selection for knowledge discovery and data mining*. Springer Science & Business Media, 1998.

[37] M. Maffina and R. RamPriya. An improved and efficient message passing interface for secure communication on distributed clusters. In *Recent Trends in Information Technology (ICRTIT), 2013 International Conference on*, pages 329–334. IEEE, 2013.

[38] M. Michael, J. E. Moreira, D. Shiloach, and R. W. Wisniewski. Scale-up x scale-out: A case study using nutch/lucene. In *Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International*, pages 1–8. IEEE, 2007.

[39] E. A. Mohammed, B. H. Far, and C. Naugler. Applications of the mapreduce programming framework to clinical big data analysis: current landscape and future trends. *BioData mining*, 7(1):22, 2014.

[40] S. Narayan, S. Bailey, and A. Daga. Hadoop acceleration in an openflow-based cluster. In *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*, pages 535–538. IEEE, 2012.

[41] N. Nurain, H. Sarwar, M. P. Sajjad, and M. Mostakim. An in-depth study of map reduce in cloud environment. In *Advanced Computer Science Applications and Technologies (ACSAT), 2012 International Conference on*, pages 263–268. IEEE, 2012.

[42] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips. Gpu computing. *Proceedings of the IEEE*, 96(5):879–899, 2008.

[43] R. S. Parpinelli and H. S. Lopes. New inspirations in swarm intelligence: a survey. *International Journal of Bio-Inspired Computation*, 3(1):1–16, 2011.

[44] R. S. Parpinelli and H. S. Lopes. Biological plausibility in optimisation: an ecosystemic view. *International Journal of Bio-Inspired Computation*, 4(6):345–358, 2012.

[45] D. Peralta, S. del Río, S. Ramírez-Gallego, I. Triguero, J. M. Benitez, and F. Herrera. Evolutionary feature selection for big data classification: A mapreduce approach. *Mathematical Problems in Engineering*, 501:246139, 2015.

[46] S. Rajaraman, G. Vaidyanathan, and A. Chokkalingam. Performance evaluation of bio-inspired optimization algorithms in resolving chromosomal occlusions. *Journal of Medical Imaging and Health Informatics*, 5(2):264–271, 2015.

[47] R. B. Sachin and M. S. Vijay. A survey and future vision of data mining in educational field. In *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on*, pages 96–100. IEEE, 2012.

[48] E. Sivaraman and R. Manickachezian. High performance and fault tolerant distributed file system for big data storage and processing using hadoop. In *Intelligent Computing Applications (ICICA), 2014 International Conference on*, pages 32–36. IEEE, 2014.

[49] C. Smutnicki. New trends in optimization. In *Intelligent Engineering Systems (INES), 2010 14th International Conference on*, pages 285–285. IEEE, 2010.

[50] R. Storn and K. Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[51] P. P. Tallon. Corporate governance of big data: Perspectives on value, risk, and cost. *Computer*, 46(6):32–38, 2013.

[52] D. Wu, Z. Li, R. Bie, and M. Zhou. Research on database massive data processing and mining method based on hadoop cloud platform. In *Identification, Information and Knowledge in the Internet of Things (IIKI), 2014 International Conference on*, pages 107–110. IEEE, 2014.

[53] J. Xie, S. Yin, X. Ruan, Z. Ding, Y. Tian, J. Majors, A. Manzanares, and X. Qin. Improving mapreduce performance through data placement in heterogeneous hadoop clusters. In *Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on*, pages 1–9. IEEE, 2010.

[54] X.-S. Yang, Z. Cui, R. Xiao, A. H. Gandomi, and M. Karamanoglu. *Swarm intelligence and bio-inspired computation: theory and applications*. Newnes, 2013.

[55] Y. Yang, X. Ni, H. Wang, and Y. Zhao. Parallel implementation of ant-based clustering algorithm based on hadoop. In *Advances in Swarm Intelligence*, pages 190–197. Springer, 2012.

[56] H. Yu, J. E. Moreira, P. Dube, I.-h. Chung, and L. Zhang. Performance studies of a websphere application, trade, in scale-out and scale-up environments. In *Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International*, pages 1–8. IEEE, 2007.

[57] I. Zelinka, D. Davendra, J. Lampinen, R. Senkerik, and M. Pluhacek. Evolutionary algorithms dynamics and its hidden complex network structures. In *Evolutionary Computation (CEC), 2014 IEEE Congress on*, pages 3246–3251. IEEE, 2014.

Paralelismo em Prolog: Conceitos e Sistemas

Parallelism in Prolog: Concepts and Systems

João Fabrício Filho [1] [2]
Anderson Faustino da Silva [2]

Data de submissão: 27/10/2015, *Data de aceite: 25/04/2016*

Resumo: Paralelismo é uma área de estudo que cresce a cada dia, devido à redução do custo e popularização de máquinas com arquiteturas paralelas. Nesse contexto, as linguagens lógicas, sobretudo o PROLOG, apresenta uma alternativa viável e prática de paralelismo. A exploração desse paralelismo pode ser realizada de diferentes formas, e há inúmeros desafios nessa tarefa. Este tutorial visa apresentar os principais conceitos de paralelismo em PROLOG, os desafios enfrentados quando se busca a paralelização nessa linguagem e o estado-da-arte do desenvolvimento de sistemas que dão suporte à paralelização em linguagens lógicas. São apresentados sistemas baseados em paralelismo implícito implementados em diferentes plataformas. Ao final é realizada uma comparação entre os sistemas apresentados e os modelos neles implementados.

***P*alavras-chave:** prolog, paralelismo, paralelismo-E, paralelismo-OU, paralelismo implícito

Abstract: Parallelism is a study area that grows up each day, caused by the cost reduction and popularizing of machines with parallels architecture. In this context, the logical languages, especially PROLOG, show a feasible and practical alternative of parallelism. This exploitation can be accomplished of different ways, and are there several challenges on this task. This survey aims to show the main concepts of parallelism in PROLOG, the faced challenges when aims to do parallelism in this language and the state-of-art of systems development to give parallelism support in logical languages. Systems with basis on implicit parallelism developed in different platforms are presented. At the end, is accomplished a comparison between the presented systems and the implemented models by they.

***K*eywords:** prolog, parallelism, AND-Parallelism, OR-Parallelism, implicit parallelism

[1]Universidade Tecnológica Federal do Paraná - Câmpus Campo Mourão Via Rosalina Maria dos Santos, 1233 CEP 87301-899 - Campo Mourão/PR - Brasil. {joaof@utfpr.edu.br}
[2]Departamento de Informática, Universidade Estadual de Maringá, Avenida Colombo, 5790 - Bloco C56 CEP 87020-900 - Maringá/PR - Brasil. {pg48335@uem.br}{anderson@din.uem.br}

1 Introdução

Os limites físicos dos componentes eletrônicos levaram ao desenvolvimento de máquinas com mais de um núcleo de processamento, e à consequente enfatização do estudo de técnicas de programação paralela, ou paralelização. A paralelização mostrou-se uma alternativa completamente viável para a maximização do processamento.

O sistema declarativo das linguagens lógicas facilita a exploração do multiprocessamento da linguagem, já que a sincronização de dados é o maior desafio ao se paralelizar um código. O paralelismo é um meio de aumentar a eficiência da execução dos códigos de programação lógica, já que se aproveita da utilização de outros núcleos de processamento, existentes em máquinas modernas.

PROLOG é a mais difundida linguagem de programação lógica e tem como princípio a cláusula de Horn [24], sendo um programa composto por regras no formato H:- $\alpha_1, \alpha_2, ..., \alpha_n$, no qual H é a cabeça do programa lógico e α_i é a iésima parte que define o corpo da cláusula. A cabeça e os alvos podem possuir zero ou mais argumentos, que podem ser termos simples (inteiros, *strings*, átomos) ou complexos (listas e estruturas).

Assim como em qualquer outro tipo de linguagem, há dois meios de paralelização de linguagens lógicas, explícita e implicitamente. Na paralelização explícita são adicionados comandos na linguagem, com os quais o usuário pode escolher quais funcionalidades serão utilizadas. Na paralelização implícita, não há diferença para o usuário em relação à programação sequencial, o interpretador procura por maneiras de paralelizar o código automaticamente, sem afetar o significado semântico.

Este tutorial tem como objetivo fazer uma apresentação da área de paralelismo em PROLOG, mostrando os principais conceitos, uma breve descrição do estado-da-arte com os principais sistemas desenvolvidos e suas funcionalidades básicas.

As principais contribuições deste trabalho são listadas a seguir:

- Apresenta conceitos básicos da área de paralelismo em PROLOG, indicado para iniciar pesquisadores nessa área;
- Lista os principais sistemas que desenvolvem paralelismo implícito ou explícito em PROLOG e suas principais funcionalidades;
- Realiza uma comparação entre esses sistemas paralelos, destacando o tipo de paralelismo implementado em cada um;
- Aborda conceitos relativamente recentes para a área, como MapReduce e *multithreading*, além dos sistemas que implementam esses conceitos;

- Realiza a conexão dos conceitos apresentados com o desenvolvimento dos sistemas, facilitando ao pesquisador a visualização das funcionalidades dos sistemas.

O restante deste tutorial está organizado da seguinte forma: a Seção 2 apresenta os conceitos básicos de paralelismo em PROLOG para melhor entendimento da área, a Seção 3 lista e discorre sobre ferramentas que implementam paralelismo em PROLOG e suas principais funcionalidades, e por fim a Seção 4 apresenta as conclusões deste trabalho.

2 Conceitos Básicos

As duas principais estratégias de exploração de paralelismo em PROLOG foram propostas por Conery e Kibler [8], e são chamadas de paralelismo-E e paralelismo-OU. Cada uma das estratégias traz estruturas de análise diferentes e modelos diferentes de implementação, os quais são abordados nas próximas seções.

Conceitos como *overhead* e granularidade estão estritamente ligados, já que definindo bem a granularidade de um sistema diminui-se o *ovehead* e ganha-se desempenho. Recentemente, abordagens com alto volume de dados são exigidas para processamento paralelo, e o modelo MapReduce [16] pode suprir essa exigência. Por último, as máquinas multinúcleo trouxeram a necessidade de suporte à paralelismo por meio do *multithreading*, que há plataformas que adotam por padrão na linguagem.

2.1 Paralelismo-E

O paralelismo-E [8] pode ocorrer quando há alguma separação na descrição de uma regra. Esse tipo de paralelismo permite a exploração da programação concorrente trazendo uma solução final para o alvo original.

Se houver uma regra definida por:

$a(X, Y)$:- $b(Y)$, $c(X)$, $d(Y, X)$.

o interpretador pode considerar o paralelismo-E no momento de executar uma consulta que leve à $a(X, Y)$ conforme o fluxo representado na Figura 1.

É comum na literatura a divisão do paralelismo-E em dois tipos: Paralelismo-E Dependente e Paralelismo-E Independente.

Ocorre o Paralelismo-E Dependente quando um ou mais alvos executados em paralelo compartilham dados entre si, esses alvos são chamados de alvos dependentes. A execução dos alvos dependentes ocorre com um dado de ligação comum entre os dois alvos, e há cooperação entre os processos diferentes. Na regra do exemplo citado anteriormente, há o

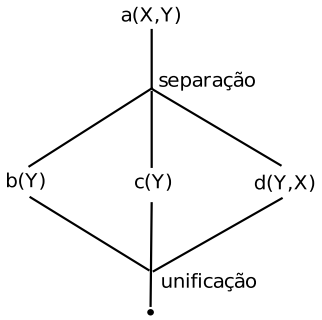


Figura 1. Fluxo de execução para a cláusula $a(X, Y)$ explorando o paralelismo-E. A separação define-se como a criação e atribuição das tarefas para diferentes fluxos de execução e a unificação como a junção dos fluxos de execução.

Paralelismo-E Dependente na execução de $b(Y)$ e $d(Y, X)$ por causa do dado Y compartilhado entre essas cláusulas.

O Paralelismo-E Independente ocorre quando os alvos paralelizados na execução não têm relações entre si, sem compartilhamento de dados. A execução desse tipo de paralelismo pode ocorrer independentemente do início ou final da execução de qualquer uma das outras regras, a junção para a solução final ocorrerá quando todos os alvos forem executados. O Paralelismo-E Independente ocorre em diversas aplicações, principalmente em problemas do tipo dividir-para-conquistar, nos quais são realizadas chamadas recursivas independentes que podem ser executadas em paralelo. Na regra introduzida no exemplo anterior, ocorre o Paralelismo-E Independente nas execuções simultâneas de $b(Y)$ e $c(X)$, que não possuem dados compartilhados entre si.

Existem alguns modelos de paralelismo na literatura que exploram o Paralelismo-E Independente. O modelo abstrato de Conery [9] foi o primeiro a ser proposto, em 1983, e explorava o Paralelismo-E Independente por meio de um grafo de fluxo de dados, no qual poderia existir uma relação produtor/consumidor entre literais se houvesse um dado não instanciado em comum. Um algoritmo de ordenamento em tempo de execução determina qual literal deverá ser executado antes e quais podem ser executados em paralelo. Há mais dois motores de execução nesse modelo além do algoritmo de ordenamento, a *execução para frente* manipula mensagens e atribui relação de produtor à literais que podem ser executados como resultado, e a *execução para trás* verifica falhas e tarefas que precisam ser refeitas. Apesar da verificação de dependência ocorrer de forma simples, há um substancial *overhead* em tempo de execução.

O modelo APEX (*And Parallel EXecution*) [28] consiste de dois algoritmos: *execução para frente* e *execução para trás*. A execução para frente pode ser vista como um esquema conceitual de passagem de *token*. A cada dado na execução do programa, um *token* é criado

e, se um literal produz um dado, ele possui o *token* desse dado. Um literal pode ser executado se ele receber os *tokens* de todos os seus dados não instanciadas do ambiente de compartilhamento atual. O algoritmo de execução para trás realiza verificação de falhas e realiza *backtracking* para refazer tarefas.

No modelo RAP (*Restricted And Parallelism* - Paralelismo-E Restrito) [18] um algoritmo de tipagem monitora os termos e predicados durante o tempo de execução do código, para manter uma indicação dos tipos de dados e construir um grafo de execução de expressões, que determina a independência entre termos e cria procedimentos de paralelização para termos independentes.

2.2 Paralelismo-OU

A máquina PROLOG possui rotinas que analisam o código antes de executá-lo. Rotinas não determinísticas são aquelas que possuem mais de uma cláusula que pode ser executada por certa sequência de argumentos. Um ponto de escolha corresponde a uma marca da busca, necessária para a restauração do estado anterior da computação. No momento em que a busca encontra uma falha e é necessário retornar, o controle passa então para o último ponto de escolha que continuará a busca pela cláusula correta. Esse retorno da busca é chamado *backtracking*, e é isso que o paralelismo-OU [8] explora na execução do código PROLOG.

Basicamente, o paralelismo-OU é possível quando existe mais de um corpo para descrever a mesma regra. O que ocorre não é paralelismo da execução, mas uma paralelização da busca pela solução por diferentes alternativas. O exemplo abaixo descreve a definição da cláusula a:

$a(X, Y)$:- $b(Y)$.
$a(X, Y)$:- $c(X)$.
$a(X, Y)$:- $d(Y, X)$.

Nesse exemplo, ao se fazer uma consulta que leve à '$a(X, Y)$', a busca pode ser realizada em paralelo efetuando as consultas por $b(Y)$, $c(X)$ e $d(Y, X)$. Por sua vez, essas consultas executam em paralelo as consultas pelas suas regras. A estrutura de uma árvore de busca é construída ao realizar o paralelismo-OU, a árvore construída no exemplo citado é representada na Figura 2. O nó raiz da árvore contém a consulta que deu origem à busca e a lista de alvos associados. Em cada um dos outros nós estão os resultados da união do primeiro alvo do nó pai, com a cabeça do programa lógico.

Diferentemente do paralelismo-E, não há junção após encontrar as definições que satisfazem a consulta realizada, o que há é a necessidade de encerrar as outras ramificações da busca, que devem ser finalizadas para evitar processamento desnecessário, já que a busca terminou.

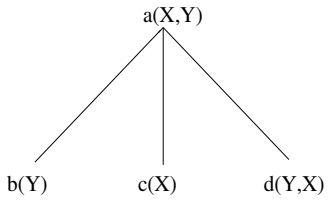


Figura 2. Árvore de busca paralela OU estruturada para a consulta de 'a.'.

Embora paralelismos E e OU sejam duas formas diferentes de explorar paralelismo em linguagens lógicas, com a exploração de ambos de forma adequada uma plataforma consegue melhor desempenho do que com a exploração de cada tipo isoladamente [29].

2.3 Efeitos colaterais

Os programas lógicos utilizam predicados chamados extra-lógicos, que não são parte da programação lógica, mas são necessários para o desenvolvimento na linguagem. São chamados de efeitos colaterais os resultados e a execução desses predicados na programação lógica. São exemplos de predicados extra-lógicos: `read`, `write`, `assert`, `retract` e `cut`.

O suporte a predicados extra-lógicos é uma das barreiras para dar suporte paralelo à linguagem PROLOG por completo. Tal suporte pode trazer vantagens, como a execução de programas existentes sem a necessidade de alteração no código. E há melhoria de desempenho no paralelismo de tais programas, que inicialmente podem ter sido desenvolvidos para ambientes sequenciais.

Dentre os predicados extra-lógicos, vale destacar a função do *corte* (ou `cut`, representado pelo símbolo !), que faz com que o programa descarte alternativas desnecessárias para uma regra. No paralelismo-OU, a execução paralela das alternativas pode fazer com que haja execução desnecessária de alternativas que já poderiam ser descartadas, o que é definido como trabalho especulativo [20].

2.4 Granularidade

A granularidade diz respeito à partilha do trabalho a ser realizado pelos processos do programa [21]. A criação e o escalonamento desses processos exige um *overhead*, e a redução desse *overhead* é o desafio ao escolher a melhor técnica de granularidade.

A granularidade é um dos maiores problemas no paralelismo implícito. Isso porque paralelizar não implica diretamente em aumento de desempenho. Há uma relação entre

overhead e fluxos de execução criados, pois a cada criação de um fluxo diferente um custo é acrescido na computação, assim deve haver uma relação em que a paralelização da tarefa compense o custo gerado pela sua criação. Portanto, deve-se encontrar a melhor relação entre o número de fluxos e custo da execução do processo para maximizar a eficiência do código.

No início da década de 1990, o trabalho de Tick *et al.* [37] atribuiu pesos quantificadores para definir a granularidade dos processos por meio da complexidade de tempo. A atribuição ocorre em tempo de compilação e por meio da conversão do código-fonte em um grafo cíclico de chamada, no qual cada nó corresponde a um procedimento e cada aresta a uma chamada de procedimento em potencial. O cálculo do peso de cada procedimento é um procedimento com recursão em calda.

O trabalho de Debray *et al.* [17] utilizou técnicas para estimar limites superiores da complexidade da computação para determinar a granularidade da paralelização. As técnicas utilizadas se basearam no tamanho dos dados e na dependência entre eles. Nesse trabalho foram calculados tamanhos relativos de variáveis compartilhadas por meio de um grafo de dependência de dados, e a informação dos tamanhos para determinar equações de representação do custo computacional dos predicados. As equações são avaliadas assim que o tamanho dos dados é conhecido, em tempo de execução. Resultados experimentais demonstram que essa técnica pode obter ganho de desempenho significativo, porém, em alguns casos pode gerar *overhead* pelas demasiadas informações processadas em tempo de execução.

Resultados mostraram que, para o paralelismo-OU, as decisões de granularidade não dependiam somente de análises de complexidade de tempo. Em 1997, o trabalho de King *et al.* [27] propôs uma técnica de análise de complexidade que englobava modelos de predicados recursivos, já que as análises anteriores não eram consideradas satisfatórias para esses tipos de problemas. Nessa técnica, considerou-se a computação em todos os tamanhos e complexidade de dados.

O trabalho de Xirogiannis [44] propôs uma técnica de análise de granularidade com foco em plataformas distribuídas com paralelismo implícito. A maior parte da análise é realizada em tempo de compilação, com alguns processamentos em tempo de execução, com aplicação experimental em paralelismo-E e paralelismo-OU.

Em 1998, o trabalho de Shen [31] tomou como base resultados que mostraram que a estimativa de complexidade não é um métrica ideal no que diz respeito à granularidade para processos criados dinamicamente, e propôs uma métrica que considera os principais *overheads* incorridos na execução do código. Nessa métrica, chamada *distance*, considera-se também os *overheads* indiretos, que são causados por tarefas paralelas que podem ser criadas dinamicamente por outras tarefas paralelas. A granularidade definida pela métrica *distance* foi experimentada em métodos executados em tempo de compilação e em tempo de execução do código PROLOG e mostrou bons resultados, especialmente no método em

tempo de execução.

2.5 Escalonamento

O escalonamento diz respeito à execução e ao gerenciamento da carga de trabalho dos processos paralelos [35]. Para se ter um ganho eficiente de desempenho, a ordem das execuções e a distribuição correta das cargas de trabalho são itens importantes.

O escalonamento pode ser dirigido por motor ou dirigido por tarefas. É dirigido por motor quando o motor de execução examina as tarefas, e dirigido por tarefas quando as tarefas examinam motores ociosos (ou menos carregados) para executá-las. Um dos primeiros escalonadores dirigidos por motor foi desenvolvido no sistema Andorra-I [14], no qual um escalonador de primeiro nível gerencia dois escalonadores de segundo nível, um responsável pelo paralelismo-E e outro responsável pelo paralelismo-OU. Os escalonadores particionam os motores em times flexíveis que distribuem o trabalho aos processos paralelos. O escalonamento dirigido por motor é utilizado no controle da execução de processos do paralelismo-OU especulativo e se mostrou eficiente nos sistemas Muse [1] e Aurora [29].

Três tipos de escalonadores implementados na literatura merecem destaque por apresentarem bons resultados: os escalonadores *Manchester*, *Bristol* e *Darma*.

O *escalonador Manchester* [6] tenta realizar uma combinação entre trabalho disponível e processos, assim que possível. Este escalonador não foi projetado para realizar trabalho especulativo eficientemente, assim não é bem encaixado no paralelismo-OU e em casos ocasionados por efeitos colaterais.

O *escalonador Bristol* [5] tenta minimizar o *overhead de sincronização* estendendo a região global do escalonador, criando várias instâncias para cada nó e é flexível para adaptar algoritmos de compartilhamento de tarefas. O escalonador foi baseado no ambiente de cópia implementado no sistema Muse [1], e foi aprimorado para o escalonamento eficiente de trabalho especulativo no paralelismo-OU, direcionado para o sistema Aurora [29].

O *Escalonador Darma* [32] também possui foco para o paralelismo especulativo, e adequado ao sistema Aurora, tentando diminuir o trabalho perdido que o sistema possuía, causado pela não poda de buscas eliminadas, como quando se tem o operador corte.

2.6 MapReduce

MapReduce [16] é um paradigma de programação projetado para processar grande volume de dados, com uma função *map* para proceder com um par chave/valor e gerar uma chave intermediária e a função *reduce* para fundir as chaves intermediárias com seus respectivos valores.

Proposto inicialmente para ser aplicado em linguagens funcionais, o modelo também pode ser adaptado à linguagens lógicas. O trabalho de Srinivasan *et al.* [33] teve como foco experimentar uma abordagem MapReduce em linguagens lógicas com paralelização de dados e tarefas e verificar sua viabilidade. Os resultados alcançados por esse trabalho evidenciaram que a abordagem é factível quando há uma quantidade considerável de dados para o problema e quando há possibilidade de diminuição de *overheads* dinâmicos.

O trabalho de Côrte-Real *et al.* [10] implementou o modelo MapReduce para PROLOG em alto nível, com o objetivo de obter uma ferramenta que distribuísse transparentemente os trabalhos não somente em grande volume de dados, mas em tarefas de processamento de dados paralelos.

Baseado na ideia do paradigma mestre-escravo, a tentativa de obter desempenho consiste em reduzir o tempo de processamento de dados distribuindo pelos recursos computacionais disponíveis, e são dispostos três níveis hierárquicos na arquitetura do trabalho de Côrte-Real *et al.* [10]: mestres globais, mestres locais e escravos. Os mestres globais controlam o fluxo de dados e o escalonamento de alto nível, enquanto os mestres locais distribuem as tarefas e controlam os acessos dos escravos, que fazem a computação propriamente dita.

Além da arquitetura, a execução do modelo MapReduce de Côrte-Real *et al.* [10] disponibiliza um modelo flexível que pode ser executado em máquinas distribuídas para grande volume de dados.

Na implementação MapReduce são adicionados predicados à linguagem PROLOG padrão, que possibilitam a utilização do paradigma pelos usuários, por meio da plataforma YAP. Foram obtidos resultados de duas implementações, memória compartilhada e passagem de mensagem, e no geral, foram obtidos *speedups* lineares.

Um modelo híbrido de MapReduce [11] permitiu a aplicação do modelo em memórias compartilhadas e distribuídas com máquinas com diferentes configurações.

2.7 Multithreading

Multithreading consiste na disponibilização de bibliotecas ou comandos-padrão na linguagem para criação e controle de fluxos paralelos explicitamente. Várias plataformas PROLOG [23, 7, 13, 26, 34, 43] disponibilizam essa funcionalidade devido tanto à popularização das máquinas paralelas quanto à disponibilização dessa funcionalidade em outras linguagens, como C e Java.

Ao contrário dos paralelismos E e OU, em *multithreading* os códigos sequenciais necessitam alterações para utilizar o paralelismo, uma vez que a funcionalidade adiciona primitivas na linguagem padrão. A mudança pode não trazer portabilidade, mas pode fazer com que os processos sejam mapeados diretamente para diferentes processadores, além de

poderem ser controlados e reutilizados.

O trabalho de Tarau [36] teve foco na construção de uma API, dentro da plataforma Bin PROLOG, na qual fluxos diferentes são criados para cada processador da máquina, otimizando a utilização do *hardware*, e desacoplando as primitivas de alto nível de *multithreading* das operações realizadas no motor de execução para adaptar os processos e fluxos ao *hardware* em questão.

O conceito de *multithreading* também é utilizado no sistema ThOr [12], em que tal conceito é aplicado em um modelo de paralelismo implícito, que é abordado com maiores detalhes na Seção 3.10.

3 Sistemas Paralelos

Vários sistemas com implementação de paralelismo exploram as linguagens lógicas. Até o final da década de 1980, a exploração do paralelismo explícito nas chamadas *linguagens de mudança persistida* recebia uma atenção considerável. Embora fossem linguagens baseadas no PROLOG, o alteravam de forma abrupta para o usuário.

Este tutorial não foca nesse tipo de linguagem, mas em ferramentas que exploram o paralelismo em PROLOG alterando-o o mínimo possível. Esta seção lista alguns sistemas e apresenta um breve resumo sobre suas funcionalidades.

3.1 PEPSys

O *Parallel ECRC PROLOG System* (PEPSys) foi apresentado em 1988 por Baron *et al.* [4], com o objetivo de avaliar novas soluções para os problemas de paralelização em programação lógica. Apesar de implementar facilidades para os paralelismos E e OU, é um sistema baseado em paralelismo explícito, no qual o controle é do usuário e são adicionados comandos à linguagem PROLOG.

Para execução do paralelismo-E, as cláusulas da regra a serem paralelizadas são separadas explicitamente por um terminal ("#"). No paralelismo-OU é necessária uma declaração de propriedades do predicado para facilitar a busca pela definição correta.

Em teoria, o paralelismo explícito dessas técnicas reduz o *overhead*, pois ninguém melhor do que o desenvolvedor para conhecer o código, escolher e saber o que é necessário paralelizar. O problema da definição da granularidade do programa assim inexiste para o sistema, ficando a critério do desenvolvedor a quantidade de computação escolhida para cada processo.

Por adicionar comandos na linguagem, programas escritos em PROLOG nativo pre-

cisam ser reescritos para utilizarem o PEPSys, o que é uma desvantagem quando se procura portabilidade e dinamicidade, pois um código escrito para uma máquina com n núcleos pode não ser tão eficiente para uma máquina com $2n$ núcleos. Além disso, existe a introdução de complexidade para o desenvolvimento de novos programas. A inexperiência do usuário também pode trazer problemas na execução do sistema, pois a granularidade do código fica a seu critério.

3.2 Aurora

Aurora [29] implementa o paralelismo-OU na linguagem PROLOG com foco em arquiteturas de memória compartilhada. O objetivo do sistema era explorar implicitamente o paralelismo-OU, para causar o mínimo de impacto possível ao usuário ao utilizar o sistema.

O sistema foi desenvolvido com base na ideia de que a exploração transparente do paralelismo-OU é mais produtiva e mais fácil de implementar do que do paralelismo-E, explorado por sistemas que implementavam o paralelismo explícito estendendo o PROLOG.

O compartilhamento de dados é feito por meio de arranjos de ligação, uma estrutura privada de cada processo, na qual os processos armazenam ligações condicionais. Para facilitar e diminuir o tempo de acesso, a estrutura é construída como um vetor compartilhado, tendo como índice um número, que representa a quantidade de variáveis criadas no ramo atual.

A exploração somente do paralelismo-OU simplifica o problema da granularidade do sistema, que pode utilizar granularidade alta, que se mostrou mais efetiva para esse tipo de paralelismo.

O compartilhamento dos dados tomou como base o modelo SRI, proposto por Warren [41], e foi construído adaptando-se à plataforma SICStus PROLOG [7].

Aurora ainda dá suporte a anotações no código-fonte em que o usuário pode optar por não paralelizar certa porção do código, fazendo com que o sistema não atue nessa porção.

O sistema conseguiu bons resultados, que evidenciaram uma redução considerável do *overhead* dos processos comparados aos sistemas PROLOG da época. Porém, a exploração somente do paralelismo-OU não traz o melhor desempenho possível que se possa ter com paralelização de PROLOG. Para obter o melhor desempenho deve-se implementar todos os tipos de paralelismo reduzindo ao máximo o *overhead* criado.

3.3 Muse

Multi Sequential PROLOG Engines (Muse) é uma abordagem apresentada em [1] que inicialmente explorou o paralelismo-OU em uma variante da linguagem PROLOG, em arqui-

teturas de memória compartilhada. Essa variante consiste na exploração do paralelismo-OU em uma extensão do SICStus PROLOG [7].

A implementação de Muse foi baseada no desenvolvimento de um sistema para paralelismo-OU chamado BC-Machine [2], focado em arquitetura de memória distribuída. A grande contribuição do sistema Muse foi o desenvolvimento do ambiente de cópia incremental, desenvolvido para minimizar o *overhead de sincronização*, fazendo com que cada instância dentro da mesma ramificação de busca possua uma parte consistente de dados e outra com modificações locais, e a sincronização dos dados é feita copiando apenas a diferença entre o local e global. Tal ambiente foi base para o escalonador Bristol [5] e para outro sistema que explora o paralelismo-OU em PROLOG, o YapOr [30].

O sistema permite ainda a utilização de anotações pelo usuário para forçar paralelismo. Tais anotações podem auxiliar no tratamento que o sistema dá ao código, porém podem deixá-lo estático e não-portável, uma vez que as anotações são específicas deste sistema.

No caso de outras funcionalidades, como corte e efeitos colaterais, o tratamento que o sistema Muse dá é desabilitar o paralelismo em seus escopos, não explorando a especulação de busca.

Muse utiliza uma abordagem em que não é necessário um coletor de lixo global, porém em casos como o da cópia incremental é preciso um processo local de coletor de lixo assíncrono.

3.4 Andorra-I

Andorra-I é um sistema proposto por Costa *et al.* [14] para explorar o paralelismo-E dependente e o paralelismo-OU, ambos de forma implícita, dando suporte à linguagem PROLOG padrão e aos programas escritos em linguagens de mudança persistida.

Este sistema implementa o modelo Andorra básico, um modelo de execução para programas lógicos que utiliza a ideia de que as cláusulas devem ser executadas assim que possível. Há dois componentes principais: o motor de execução e o preprocessador determinístico. Enquanto o motor de execução interpreta os comandos na linguagem-fonte, o preprocessador verifica possíveis ramificações para uma busca paralela.

A computação determinística pode ter vários predicados executados em paralelo, compondo o paralelismo-E independente. Quando não há alvo determinado (computação não-determinística) um alvo é escolhido e um ponto de escolha é criado para o alvo, e os vários pontos de escolha podem ser executados em paralelo, compondo o paralelismo-OU.

Quanto ao escalonamento implementado, o escalonador Bristol [5], que foi desenvolvido para o sistema Aurora, foi adaptado ao sistema Andorra-I.

3.5 &-Prolog

O sistema &-Prolog foi apresentado em 1991 por Hermenegildo e Greene [22] com o objetivo de prover um ambiente no qual o usuário pode explorar a execução paralela de programas lógicos automaticamente, ou adicionando anotações de divisão de tarefas entre fluxos.

A implementação se dá por meio do paralelismo implícito, com código PROLOG nativo, e explícito, no qual a adição de terminais ao código-fonte permite ao usuário explorar o paralelismo-E.

A análise do código em tempo de compilação se dá pelo Grafo de Expressão Condicional (GEC), que possui a forma ($i\_cond \ => \ goal_1 \ \& \ goal_2 \ \& \ ... \ \& \ goal_n$), na qual i_cond representa uma condição que, se satisfeita, permite a independência de $goal_i$, para todo $i < n$.

O sistema &-Prolog funciona basicamente com quatro módulos: anotador, analisador de efeitos colaterais, analisador global e compilador de baixo nível.

O anotador realiza a análise de dependência no código de entrada, verificando quais tarefas podem ser executadas em paralelo. O analisador de efeitos colaterais verifica se o predicado atual contém ou chama efeitos colaterais. O analisado global interpreta o código no domínio abstrato do sistema e infere sobre possíveis substituições ao longo do programa. Por fim, o compilador de baixo nível funciona como uma extensão do SICStus PROLOG, produzindo código intermediário do programa de entrada.

3.6 Opera

O projeto Opera, apresentado no trabalho de Werner *et al.* [42], explora o paralelismo-E restrito e o paralelismo-OU multi-sequencial de forma implícita, apesar de disponibilizar minimamente primitivas opcionais ao usuário para paralelização.

O paralelismo-OU multi-sequencial é implementado sem compartilhamento de variáveis, caso duas ou mais operações precisem do mesmo dado, o dado é copiado para uma nova instância na memória. O autor explorou tal alternativa visando arquiteturas de memória compartilhada, nas quais o acesso à mesma região de memória de diferentes processos pode ser computacionalmente custoso. O paralelismo-E restrito é empregado por meio de análises de estruturas em tempo de compilação.

O sistema constrói uma estrutura de análise de granularidade, na qual busca a máxima granularidade, utilizando uma heurística de que, quanto mais alto nível for o trabalho, maior sua granularidade. O *overhead* de criação de um processo aumenta quando considera-se que os dados são copiados a cada criação.

O foco na arquitetura de memória distribuída com paralelismos E e OU traz uma vantagem ao sistema Opera. A opção de primitivas para paralelismo explícito é um atrativo para os usuários. Porém, tais primitivas podem deixar o código estático e não portável.

3.7 YapOr

YapOr é um sistema originalmente publicado por Rocha *et al.* [30] que explora o paralelismo-OU em uma extensão do ambiente YAP PROLOG [13].

O ambiente YAP PROLOG é baseado na WAM [40], e portanto possui várias otimizações que não são presentes na versão original do SICStus PROLOG, ambiente dos sistemas Muse e Aurora. Assim, um sistema com base na plataforma YAP tem mais condições de ganho de desempenho do que em sistemas com base em plataformas que não possuam essas otimizações.

A implementação do paralelismo-OU no YapOr é baseada no ambiente de cópia utilizado no sistema Muse, que, segundo os autores, é mais simples, elegante e adequado à complexidade da plataforma YAP. Porém, a organização da memória e a implementação de travas são as diferenças entre o desenvolvimento de YapOr e Muse.

A comparação com Muse é inevitável, visto que o desenvolvimento de YapOr foi baseado em Muse. No trabalho em que o YapOr foi apresentado [30], há uma seção em que foram comparados resultados de execução de um mesmo conjunto de programas nos dois ambientes. Apesar de na época o ambiente Muse estar mais maduro, YapOr obteve melhor desempenho.

A implementação do corte no YapOr é realizada por meio da execução especulativa, o envio de sinais pelas raízes de busca para os nós finaliza a busca. A busca especulativa se inicia ao encontrar as alternativas, mas ao receber o sinal os nós são podados. Tal implementação minimiza o *overhead* na busca, mas por outro lado, ainda inicia busca desnecessárias.

3.8 PALS

PALS [39] é um sistema publicado em 2001 que explora o paralelismo-OU em arquiteturas de memória distribuída, construído com base no ambiente de cópia presente em sistemas como Muse [1] e YapOr [30].

Além do ambiente de cópia, a cópia incremental também presente nos sistemas citados foi implementada no sistema PALS, se mostrando eficiente para exploração do paralelismo-OU. Resultados da execução do sistema, que tem como base a plataforma ALS PROLOG [3], mostraram que o ambiente de cópia é adequado à arquitetura de memória distribuída, reduzindo *overheads* significativamente. Tal redução pode ser causada pela minimização das

atualizações de memórias, que possuem custo significativo nessa arquitetura.

No que diz respeito ao escalonamento, o sistema utiliza uma técnica chamada de divisão de pilha, tida como uma evolução das técnicas utilizadas no ambiente de cópia de Muse e YapOr. Na divisão de pilha, a distribuição de tarefas se realiza no momento da separação, evitando estruturas de dados centralizadas para escalonar.

3.9 PAN

PAN [45] é um ambiente distribuído de propósito geral e portável para executar programas lógicos em paralelo, explorando as paralelizações dos tipos dividir-e-conquistar e especulativa.

Originalmente publicado em 2002, o sistema combina SICStus PROLOG [7] e PVM [15] para gerar uma arquitetura distribuída. A exploração do paralelismo se dá por meio de análises em tempo de compilação, sem interferência do usuário (paralelismo implícito). De um modo grosseiro, o sistema PAN cria processos estaticamente, que se comunicam por meio de mensagens síncronas e assíncronas, utilizando primitivas adicionadas à linguagem PROLOG para controlar a passagem de mensagem e habilitar a sincronização.

O sistema PAN dá suporte ao paralelismo-OU e ao paralelismo-E independente, e, apesar de a comunicação da arquitetura exigir adição de primitivas, o paralelismo se dá de forma automática, com a linguagem PROLOG padrão.

3.10 ThOr

Com o crescimento e popularização das máquinas multinúcleos, houve necessidade de adequar o paralelismo implícito à melhor distribuição dos fluxos entre os núcleos disponíveis.

Threads and Or-Parallelism Unified (ThOr) [12] é uma extensão do sistema YapOr, com o objetivo de integrar o paralelismo-OU com a biblioteca de *multithreading* da plataforma YAP. A biblioteca de *multithreading* do YAP consiste basicamente em uma interface alto nível para POSIX Threads [25].

A exploração de *multithreading* junto com o paralelismo-OU permite maior escalabilidade do sistema, com uma distribuição mais uniforme das tarefas pelos núcleos de processadores disponíveis na máquina.

Melhorias foram realizadas no sistema recentemente [19] e, no geral, conseguiram resultados mais eficientes do que YapOr.

3.11 Uma comparação

Os sistemas apresentados exploram os paralelismos E e OU de forma implícita em sua maioria, isso permite que o usuário não tenha preocupação quanto às oportunidades de paralelismo do programa, o sistema de execução paraleliza seu programa automaticamente.

Alguns sistemas [1, 29, 42] disponibilizam alternativas de anotações adicionais à linguagem padrão, apesar de denominados sistemas de paralelismo implícito. Tais anotações podem disponibilizar mais funcionalidades ao usuário, porém o código pode ficar estático e não-portável.

O paralelismo-E independente e modelos que exploram esse paralelismo são os mais implementados devido à facilidade que proporcionam na interpretação do código.

A arquitetura de memória compartilhada é a mais explorada na área de paralelismo em PROLOG, porém há sistemas que disponibilizam essa alternativa nas arquiteturas de memória distribuída.

A Tabela 1 apresenta uma comparação dos sistemas apresentados nesta seção. Na coluna 'arquitetura' entede-se por 'MD' arquiteturas de memória distribuída e 'MC' arquiteturas de memória compartilhada.

Tabela 1. Síntese dos sistemas apresentados que implementam paralelismo em PROLOG

sistema	ano	plataforma	paralelismo	implementação	arquitetura
PEPSys	1988	CProlog	explícito	E e OU	MD
Aurora	1990	SICStus	implícito*	OU	MC
Muse	1990	SICStus	implícito*	OU	MC
Andorra-I	1991	Andorra	ambos	E e OU	MC
&-Prolog	1991	Quintus	ambos	RAP	MC
Opera	1992	MAP	implícito*	RAP e OU	MC
YapOr	1999	YAP	ambos	OU	MC
PALS	2001	ALS	implícito	OU	MD
PAN	2002	SICStus/PVM	implícito	E e OU	MD
ThOr	2010	YAP	ambos	OU	MC

* com suporte a anotações

O paralelismo-OU é mais explorado que o paralelismo-E na literatura, pois os resultados apresentados apresentam melhor desempenho em relação a seus custos. No contexto de

paralelismo-OU em arquiteturas de memória compartilhada, YapOr e Muse apresentaram os melhores resultados [38].

A maximização da exploração do paralelismo ocorre somente quando explorados ambos os tipos de paralelismo, mas apesar disso, diversos sistemas [30, 1, 29, 39] apresentam bons resultados explorando apenas um tipo de paralelismo.

O ambiente de cópia implementado inicialmente no sistema Muse, adaptado aos sistemas YapOr e PALS, foi uma grande contribuição e permitiu a redução principalmente do *overhead* de sincronização entre processos.

Atualmente, com a popularização de máquinas com arquitetura de vários núcleos, o *multithreading* tornou-se necessário em plataformas que necessitam expandir seus ganhos de desempenho nessas máquinas, havendo sistemas que implementaram esse tipo de solução [12].

4 Conclusão

O paralelismo está presente na computação como um paradigma de implementação, e é uma área com frutíferos estudos. Nesse contexto, as linguagens lógicas têm mostrado um gancho potencial para a exploração de paralelismo.

A exploração do paralelismo implícito em PROLOG mostra uma maneira de fazer paralelismo sem alterações para o usuário, além de dar suporte a todas as aplicações pré-existentes na linguagem.

Vários conceitos foram abordados neste tutorial, que exemplificam a complexidade da área de estudo e dos problemas enfrentados. A minimização do *overhead* ao se paralelizar as aplicações PROLOG pode ser realizada com diferentes estratégias, dependendo do tipo de aplicação e de paralelismo desenvolvido.

Uma análise dos sistemas atuais que exploram o paralelismo indica que há sistemas consolidados que exploram um tipo de paralelismo (caso de YapOr, Muse e Aurora, que exploram somente o paralelismo-OU) ou em certa arquitetura (PAN e Opera, que executam em arquiteturas de memória distribuída e compartilhada, respectivamente).

Com a popularização de máquinas com vários núcleos computacionais, PROLOG é uma alternativa elegante para a exploração de tal paralelismo.

Contribuição dos autores:

Os dois autores contribuíram igualmente para a elaboração de todas as seções.

Referências

[1] ALI, K., AND KARLSSON, R. The muse approach to or-parallel prolog. *International Journal of Parallel Programming 19*, 2 (1990), 129–162.

[2] ALI, K. M. Or parallel execution of prolog on bc-machine. *SICS Research Report* (1988).

[3] APPLIED LOGIC SYSTEMS INC. Als prolog manual. *Cambridge, MA* (1997).

[4] BARON, U., DE KERGOMMEAUX, J. C., HAILPERIN, M., RATCLIFFE, M., ROBERT, P., SYRE, J.-C., AND WESTPHAL, H. The parallel ecrc prolog system pepsys: An overview and evaluation results. In *FGCS* (1988), pp. 841–850.

[5] BEAUMONT, A., RAMAN, S. M., SZEREDI, P., AND WARREN, D. H. D. Flexible scheduling of or-parallelism is aurora: The bristol scheduler. In *Proceedings on Parallel Architectures and Languages Europe : Volume II: Parallel Languages: Volume II: Parallel Languages* (New York, NY, USA, 1991), PARLE '91, Springer-Verlag New York, Inc., pp. 403–420.

[6] CALDERWOOD, A., AND SZEREDI, P. Scheduling Or-parallelism in Aurora: The Manchester Scheduler. In *International Conference on Logic Programming/Joint International Conference and Symposium on Logic Programming* (1989), pp. 419–435.

[7] CARLSSON, M., AND MILDNER, P. Sicstus prologthe first 25 years. *Theory and Practice of Logic Programming 12* (1 2012), 35–66.

[8] CONERY, J. S., AND KIBLER, D. F. Parallel interpretation of logic programs. In *Proceedings of the 1981 Conference on Functional Programming Languages and Computer Architecture* (New York, NY, USA, 1981), FPCA '81, ACM, pp. 163–170.

[9] CONERY, J. S., AND KIBLER, D. F. And parallelism in logic programs. In *In Proceedings of the International Joint Conference on AI* (Los Altos, CA, 1983), A. Bundy, Ed., pp. 539–543.

[10] CÔRTE-REAL, J., DUTRA, I., AND ROCHA, R. Prolog programming with a mapreduce parallel construct. In *Proceedings of the 15th Symposium on Principles and Practice of Declarative Programming* (New York, NY, USA, 2013), PPDP '13, ACM, pp. 285–296.

[11] CÔRTE-REAL, J., DUTRA, I., AND ROCHA, R. A hybrid mapreduce model for prolog. In *Integrated Circuits (ISIC), 2014 14th International Symposium on* (Dec 2014), pp. 340–343.

[12] COSTA, V. S., DUTRA, I., AND ROCHA, R. Threads and or-parallelism unified. *Theory and Practice of Logic Programming 10* (7 2010), 417–432.

[13] COSTA, V. S., ROCHA, R., AND DAMAS, L. The yap prolog system. *Theory and Practice of Logic Programming 12* (1 2012), 5–34.

[14] COSTA, V. S., WARREN, D. H. D., AND YANG, R. Andorra i: A parallel prolog system that transparently exploits both and-and or-parallelism. *SIGPLAN Not. 26*, 7 (Apr. 1991), 83–93.

[15] CUNHA, J. C., AND MARQUES, R. F. Pvm-prolog: A prolog interface to pvm. In *In Proceedings of the 1st Austrian-Hungarian Workshop on Distributed and Parallel Systems, DAPSYS'96* (1996), pp. 173–181.

[16] DEAN, J., AND GHEMAWAT, S. Mapreduce: Simplified data processing on large clusters. *Commun. ACM 51*, 1 (Jan. 2008), 107–113.

[17] DEBRAY, S. K., LIN, N.-W., AND HERMNEGILDO, M. Task granularity analysis in logic programs. In *Proceedings of the ACM SIGPLAN 1990 Conference on Programming Language Design and Implementation* (New York, NY, USA, 1990), PLDI '90, ACM, pp. 174–188.

[18] DEGROOT, D. Restricted and-parallelism and side effects. In *Procs. of the 1987 Int'l Symp. on Logic Programming (ISLP 87)* (1987), IEEE Computer Society, pp. 80–89.

[19] DUTRA, I., ROCHA, R., COSTA, V., SILVA, F., AND SANTOS, J. Scheduling or-parallelism in yapor and thor on multi-core machines. In *Parallel and Distributed Processing Symposium Workshops PhD Forum (IPDPSW), 2012 IEEE 26th International* (May 2012), pp. 1581–1590.

[20] GUPTA, G., AND COSTA, V. S. Cuts and side-effects in and-or parallel prolog. *The Journal of Logic Programming 27*, 1 (1996), 45 – 71.

[21] GUPTA, G., PONTELLI, E., ALI, K. A., CARLSSON, M., AND HERMENEGILDO, M. V. Parallel execution of prolog programs: A survey. *ACM Trans. Program. Lang. Syst. 23*, 4 (July 2001), 472–602.

[22] HERMENEGILDO, M., AND GREENE, K. The &-prolog system: Exploiting independent and-parallelism. *New Generation Computing 9*, 3-4 (1991), 233–256.

[23] HERMENEGILDO, M. V., BUENO, F., CARRO, M., LÓPEZ-GARCÍA, P., MERA, E., MORALES, J. F., AND PUEBLA, G. An overview of ciao and its design philosophy. *Theory and Practice of Logic Programming 12* (1 2012), 219–252.

[24] HORN, A. On sentences which are true of direct unions of algebras. *The Journal of Symbolic Logic 16* (3 1951), 14–21.

[25] IEEE PORTABLE APPLICATIONS STANDARDS COMMITTEE AND OTHERS. Ieee std 1003.1 c-1995, threads extensions, 1995.

[26] INTELLIGENT SYSTEMS LABORATORY. *Quintus Prolog user's manual*. Swedish Institute of Computer Science, Kista, Sweden, December 2003.

[27] KING, A., SHEN, K., AND BENOY, F. Lower-bound time-complexity analysis of logic programs. In *ILPS* (1997), vol. 97, pp. 261–275.

[28] LIN, Y.-J., AND KUMAR, V. And-parallel execution of logic programs on a shared-memory multiprocessor. *J. Log. Program. 10*, 2 (Jan. 1991), 155–178.

[29] LUSK, E., BUTLER, R., DISZ, T., OLSON, R., OVERBEEK, R., STEVENS, R., WARREN, D., CALDERWOOD, A., SZEREDI, P., HARIDI, S., BRAND, P., CARLSSON, M., CIEPIELEWSKI, A., AND HAUSMAN, B. The aurora or-parallel prolog system. *New Generation Computing 7*, 2-3 (1990), 243–271.

[30] ROCHA, R., SILVA, F., AND COSTA, V. S. YapOr: an Or-Parallel Prolog System Based on Environment Copying. In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence (EPIA 1999)* (Évora, Portugal, September 1999), P. Barahona and J. Alferes, Eds., no. 1695 in LNAI, Springer, pp. 178–192.

[31] SHEN, K., COSTA, V. S., AND KING, A. Distance: A new metric for controlling granularity for parallel execution. In *Proceedings of the 1998 Joint International Conference and Symposium on Logic Programming* (Cambridge, MA, USA, 1998), JICSLP'98, MIT Press, pp. 85–99.

[32] SINDAHA, R. The dharma scheduler-definitive scheduling in aurora on multiprocessors architecture. In *Parallel and Distributed Processing, 1992. Proceedings of the Fourth IEEE Symposium on* (Dec 1992), pp. 296–303.

[33] SRINIVASAN, A., FARUQUIE, T., AND JOSHI, S. Data and task parallelism in ilp using mapreduce. *Machine Learning 86*, 1 (2012), 141–168.

[34] SWIFT, T., AND WARREN, D. S. Xsb: Extending prolog with tabled logic programming. *Theory and Practice of Logic Programming 12* (1 2012), 157–187.

[35] TANENBAUM, A. S. *Modern Operating Systems*, 3rd ed. Prentice Hall Press, Upper Saddle River, NJ, USA, 2007.

[36] TARAU, P. Concurrent programming constructs in multi-engine prolog: Parallelism just for the cores (and not more!). In *Proceedings of the Sixth Workshop on Declarative Aspects of Multicore Programming* (New York, NY, USA, 2011), DAMP '11, ACM, pp. 55–64.

[37] TICK, E. Compile-time granularity analysis for parallel logic programming languages. *New Generation Computing 7*, 2-3 (1990), 325–337.

[38] VIEIRA, R., ROCHA, R., AND SILVA, F. Or-Parallel Prolog Execution on Multicores Based on Stack Splitting. In *Proceedings of the 7th International Workshop on Declarative Aspects and Applications of Multicore Programming (DAMP 2012)* (Philadelphia, Pennsylvania, USA, January 2012), V. S. Costa, Ed., ACM Digital Library, pp. 1–9.

[39] VILLAVERDE, K., PONTELLI, E., GUO, H., AND GUPTA, G. Pals: An or-parallel implementation of prolog on beowulf architectures. In *Logic Programming*, P. Codognet, Ed., vol. 2237 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2001, pp. 27–42.

[40] WARREN, D. H. D. An abstract prolog instruction set. Tech. Rep. 309, AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, Oct 1983.

[41] WARREN, D. H. D. The sri model for or-parallel execution of prolog: Abstract design and implementation issues. In *inProceedings of the 1987 Symposium on Logic Programming* (1987), pp. 92–102.

[42] WERNER, O., YAMIN, A. C., BARBOSA, J. L. V., AND GEYER, C. F. R. Opera project: An approach towards parallelism exploitation on logic programming. In *WLP* (1994), pp. 20–23.

[43] WIELEMAKER, J., SCHRIJVERS, T., TRISKA, M., AND LAGER, T. SWI-Prolog. *Theory and Practice of Logic Programming 12*, 1-2 (2012), 67–96.

[44] XIROGIANNIS, G. Granularity control for distributed execution of logic programs. In *Distributed Computing Systems, 1998. Proceedings. 18th International Conference on* (1998), IEEE, pp. 230–237.

[45] XIROGIANNIS, G., AND TAYLOR, H. Pan: A portable, parallel prolog: Its design, realisation and performance. *New Generation Computing 20*, 4 (2002), 373–399.

Detecção de Tubos em Imagens Radiográficas Digitais

Pipe detection in digital radiographic images

Márlon de Oliveira Vaz [1]
Tania Mezzadri Centeno [2]
Myriam Regattieri Delgado [3]

Data de submissão: 01/06/2015, *Data de aceite: 15/02/2016*

Resumo: Este artigo apresenta uma metodologia para a detecção do tubo em imagens radiográficas do tipo parede dupla vista dupla (PDVD) de tubulações condutoras de petróleo. O principal objetivo da proposta é reduzir a região de busca através da delimitação da área do tubo para a extração automática do cordão de solda auxiliando, desta forma, a posterior detecção de defeitos em juntas soldadas. O processo de detecção do tubo apresentado é totalmente automático e baseado em técnicas de processamento de imagens como ajustes no brilho e contraste, limiarização e análise das regiões identificadas para a segmentação do tubo. O processo foi aplicado em 117 imagens provenientes de três diferentes sistemas radiográficos obtendo um resultado de 99,14% de acerto na detecção do tubo. Foi realizada uma comparação com outra abordagem para a detecção do tubo em imagens radiográficas do tipo PDVD e a metodologia proposta apresentou melhora em relação ao trabalho anterior. Conclui-se, portanto que o método proposto pode ser usado como uma etapa que precede a detecção automática do cordão de solda.

***P*alavras-chave:** detecção de tubos, images radiográficas, PDVD, processamento digital de imagens

[1]Doutorando em Engenharia Elétrica e Informática Industrial / UTFPR, docente Instituto Federal do Paraná - IFPR - Curitiba - PR, Brasil. {marlonvaz@gmail.com}
[2]Doutorado em Doctorat En Informatique, docente Universidade Tecnológica Federal do Paraná - UTFPR - Curitiba-PR, Brasil. {mezzadri@utfpr.edu.br}
[3]Doutorado em Engenharia Elétrica pela Universidade Estadual de Campinas, docente Universidade Federal do Paraná - UTFPR - Curitiba-PR, Brasil. {myriamdelg@gmail.com}

Abstract: This paper presents a methodology to detect pipes in double wall double image (DWDI) radiographic images of petroleum pipelines. The main goal is to reduce the search region in automatic weld beads extractions (by delimiting the pipeline area), contributing this way to the subsequently detection of defects in welded joints. The pipe detection process is fully automatic and based on image processing techniques such as brightness and contrast adjustment, thresholding and analysis of the identified regions for pipe segmentation. Experiments consider 117 images acquired from three different radiographic systems with a result of 99,14% accuracy in the pipe detection. We also compare the detection results with those provided by a previous approach (developed to detect pipes in DWDI radiographic images) and the proposed methodology provides an improvement over previous work. We can conclude that the proposed method is a potential pre-processing step for detecting weld beads.

***K*eywords:** pipe detection, radiographic images, DWDI, digital image processing

1 Introdução

Os ensaios não destrutivos (END) são amplamente utilizados pela indústria petroquímica para inspecionar soldas em tubos condutores de fluidos. Esta técnica permite ao inspetor avaliar a qualidade do cordão de solda sem causar impactos no sistema. Dentre as técnicas (END) mais utilizadas para análise da qualidade de um cordão de solda está o ensaio radiográfico. Esta técnica possibilita que uma imagem radiográfica seja gerada e posteriormente avaliada por um inspetor laudista qualificado. Além da vantagem do ensaio radiográfico poder ser utilizado na maioria das situações, este permite um registro permanente do exame realizado. Porém, a interpretação radiográfica não é tarefa simples, visto que o inspetor laudista é responsável por analisar um grande volume de imagens. Por ser uma tarefa exaustiva, os resultados podem variar quando se comparam laudos realizados por inspetores diferentes ou, até mesmo, no caso de laudos realizados pelo mesmo profissional sob diferentes circunstâncias no momento da execução da tarefa [1],[2].

Diante desse contexto e com base no desenvolvimento e aperfeiçoamento de equipamentos de radiografia digital, há atualmente uma grande quantidade de pesquisas em busca da automação parcial ou completa da interpretação radiográfica. Centros de pesquisa em diversos países têm focalizado seus esforços na tentativa de encontrar ferramentas computacionais automáticas que possam auxiliar na redução da subjetividade na interpretação radiográfica [3],[4],[5],[6],[7].

Uma área que tem avançado bastante é o estudo de técnicas de reconhecimento de padrões para classificação automática de defeitos de soldagem. Contudo, geralmente se recorre à extração manual ou semiautomática de regiões da imagem, denominadas regiões de inte-

resse (ROI do inglês regions of interest), que no caso da aplicação considerada neste artigo, incluem o cordão de solda e os defeitos a serem analisados [5],[6],[7]. Para melhorar a eficiência dos métodos de detecção de defeitos seria conveniente utilizar métodos para segmentar a região do cordão de solda a ser inspecionado de forma a reduzir o espaço de busca. A detecção prévia do tubo evita que a etapa de busca pelo cordão de solda encontre outras regiões com características semelhantes às do cordão de solda. Além disso, o tempo computacional na busca pelo cordão é reduzido. Considerando os aspectos citados, surge a necessidade de se aprimorar as técnicas de segmentação de imagens radiográficas de juntas soldadas.

Assim, a detecção do cordão de solda é um passo fundamental para a automação de processos de inspeção por meio de imagens radiográficas. Na literatura encontram-se alguns trabalhos que tratam desse processo e que utilizam diferentes técnicas obtendo resultados diversos [2],[4],[8],[9],[10],[11], [12]. No entanto, somente os trabalhos de [4],[12],[13] e [14] tratam de cordões de solda do tipo parede dupla vista dupla (PDVD) que constituem a maioria das imagens encontradas no processo de inspeção de defeitos em cordões de solda com diâmetro igual ou inferior a 3¡ polegadas. Em [4] e [12] a abordagem proposta utiliza uma detecção semiautomática do tubo, como etapa preliminar do processo de detecção do cordão de solda o que auxilia na identificação da região do cordão de solda.

Este artigo propõe um método automático para segmentar a região do tubo empregando técnicas de processamento digital de imagens, de forma que este possa ser utilizado como uma etapa anterior à detecção do cordão de solda. O trabalho apresenta, na seção 2, o referencial teórico com conceitos básicos e trabalhos correlatos. A metodologia utilizada para resolver o problema de segmentação do tubo é detalhada na seção 3. Na seção 4 são apresentados os resultados obtidos com o método proposto e a seção 5 traz as conclusões do trabalho.

2 Fundamentação Teórica

Esta seção traz os conceitos básicos necessários para o entendimento da proposta. Na seção 2.1 a técnica de exposição radiográfica PDVD é apresentada de forma sucinta. A seção 2.2 discute as técnicas de processamento de imagens que serviram de base para a metodologia proposta. Por fim, na seção 2.3, alguns trabalhos correlatos são apresentados.

2.1 Técnica de exposição radiográfica PDVD

A técnica de exposição radiográfica utilizada para a aquisição da imagem de cordões de solda está relacionada, principalmente com o tipo de montagem da tubulação e sua espessura. As tubulações com diâmetro igual ou inferior a 3¡ polegadas (90 mm) não possibilitam acesso interno para a inserção da fonte de radiação, neste caso, a imagem radiográfica é usu-

almente obtida através da técnica PDVD [3],[15].

Conforme ilustrado na figura 1, a técnica PDVD consiste em posicionar a fonte de radiação e o filme radiográfico no exterior do tubo, de forma que a radiação atravesse as duas paredes do tubo (parede dupla) e que as imagens das duas extremidades do cordão de solda sejam formadas no filme radiográfico (vista dupla). O cordão de solda é projetado no filme radiográfico apresentando um aspecto elíptico [15] A imagem da região do cordão de solda mais distante da fonte de radiação (parte superior do anel elíptico) se apresenta mais atenuada e também mais ruidosa, visto que a radiação atravessa as duas paredes da tubulação. Portanto, as abordagens usadas para tratar as imagens radiográficas provenientes dessa técnica devem considerar a falta de qualidade da imagem.

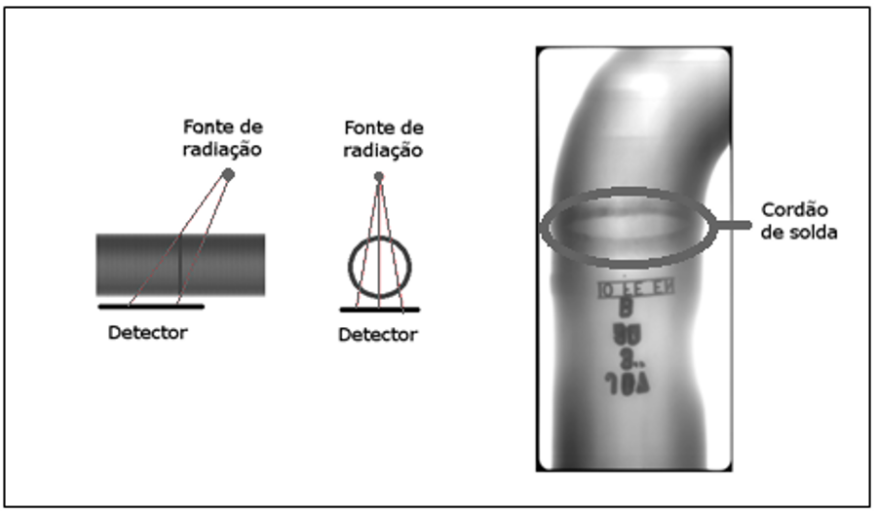


Figura 1. Técnica de exposição radiográfica PDVD com a imagem projetada do cordão de solda.

2.2 Técnicas de processamento de imagens

A região do cordão de solda a ser inspecionada (região delimitada pela elipse cinza escuro na Figura 1) constitui apenas uma pequena porção da imagem adquirida, na qual, além do tubo com marcações, aparece também a região do fundo. Desse modo, a localização prévia do cordão de solda é desejável na detecção automática de defeitos para otimizar o processo, limitando a região de busca. Entretanto, a localização automática do cordão não é trivial devido aos ruídos e pouco contraste nas bordas. Por isso a segmentação da região do tubo na qual o cordão está presente pode facilitar a busca. Diferentes técnicas de processamento digital de imagens podem ser empregadas para aprimorar a imagem e realçar as áreas de interesse facilitando a segmentação da região do tubo.

Ajuste do contraste. Um dos problemas mais frequentemente encontrados em imagens digitais é o contraste e/ou brilho inadequado o que impede que as operações de segmentação sejam realizadas de forma eficiente. As técnicas de processamento de imagens usadas para realizar o ajuste do contraste são as transformações de intensidade [16]. Quando se deseja realçar uma faixa de níveis de cinza de forma a destacar o objeto de interesse na imagem uma transformação logarítmica pode ser empregada. A forma geral da curva logarítmica dessa transformação é definida por (1) [16]:

$$s = c * log(1 + r) \tag{1}$$

onde *r* é o valor de nível de cinza do pixel atual $(r \geq 0)$, *s* é o valor de nível de cinza do pixel mapeado e *c* é uma constante escalar. A faixa dinâmica dependerá da intensidade da imagem podendo ser de $[0, 2^8 - 1]$ ou de $[0, 2^{16} - 1]$. A transformação logarítmica é usada para expandir os valores de pixels mais baixos (mais escuros) em uma imagem e simultaneamente comprimir os valores de nível mais alto (mais claros).

Limiarização. Uma das formas mais simples de segmentar uma imagem é através da limiarização que transforma uma imagem em níveis de cinza em uma imagem binária com apenas dois níveis de cinza 0 (preto) e 1 (branco). Quando regiões de uma imagem (fundo e objetos) diferem entre si pela luminosidade (valores de nível de cinza), pixels pertencentes a diferentes regiões estarão em faixas limitadas por níveis de cinza distintos. Se estas faixas forem bem definidas é possível segmentar as regiões de interesse (objetos) através da limiarização [16]. A imagem segmentada g(x, y) é obtida pela equação (2):

$$g(x, y) = \begin{cases} 1 & se \quad f(x) > T \\ 0 & se \quad f(x) \leq T \end{cases} \tag{2}$$

onde *f(x, y)* corresponde ao nível de cinza do pixel da imagem de entrada localizado na coordenada (*x*, *y*) e *T* é um limiar, valor de nível de cinza que separa os pixels do objeto e do fundo. Este processo consiste em comparar os níveis de cinza dos pixels de uma imagem ao limiar *T* e então classificá-los. Então qualquer ponto (*x*, *y*) na imagem em que *f(x, y)* > *T* é chamado de ponto do objeto; caso contrário, o ponto é chamado de ponto de fundo.

Na maioria das aplicações é importante que a determinação do limiar seja automática e para tal existem diversas técnicas. Sezgin e Sankur [17] avaliam 40 diferentes técnicas de limiarização aplicadas usualmente na área de END. Um dos métodos mais conhecidos de limiarização é o método de Otsu [18] que calcula o valor do limiar com base nas propriedades estatísticas da imagem.

Etiquetagem (*Labeling*). Na identificação de regiões o propósito é agrupar os pixels que possuem características semelhantes, como o nível de intensidade, criando assim, um conjunto de regiões [19]. Este processo pode ser aplicado para imagens binárias e imagens

em nível de cinza. O algoritmo de etiquetagem atua em cada pixel da imagem verificado a sua vizinhança, caso ocorram valores de intensidade de similares, estes são conectados com o pixel inicial. Esta relação entre o ponto de origem e seus vizinhos é definida pelo tipo de conectividade. Neste caso, há dois modos de conectividade, a vizinhança-4 (*four-neighbor*) e a vizinhança-8 (*eight-neighbor*) [20]. Ao término do algoritmo cada agrupamento identificado é marcado numericamente possibilitando assim, definir as regiões dentro da imagem.

Neste trabalho foi utilizada a função *bwlabel* que faz parte do *Image Processing Toolbox* do *Matlab*®. Esta função foi desenvolvida com base no algoritmo *run-length encoding* [21]. O algoritmo percorre uma imagem binária linha a linha a procura pixels brancos, onde P é a quantidade de conjuntos de pixels contíguos encontrados em cada linha. Ao estabelecer uma sequência de pixels esta é armazenada em uma tabela contendo a linha, o pixel de início, o pixel de fim e a classificação (*P*). Na execução da próxima linha são avaliados os pixels em relação à linha anterior determinado assim as adjacências entre os conjunto de pixels. Ao determinar a adjacência entre os grupos de pixels o algoritmo relaciona os conjuntos a uma tabela que expressa uma classe. Ao chegar à última linha o algoritmo organizou todos os conjuntos em classes.

Como já observado anteriormente, grande parte das pesquisas para a detecção de defeitos em cordão de solda realiza a extração manual da região de interesse, além disso, há poucas pesquisas que se concentram na técnica PDVD. Como será visto na próxima seção, nos trabalhos [4] e [12] a detecção do cordão de solda é realizada em dois passos: primeiramente é feita a detecção do tubo seguida da busca pelo cordão de solda no interior da imagem do tubo detectada. Entretanto, quando a imagem do tubo não é detectada, não é possível detectar o cordão de solda. Assim, este trabalho pretende contribuir nesta direção, propondo uma forma automática de detecção do tubo.

2.3 Detecção do tubo/cordão de dolda em radiografias PDVD

Como comentado anteriormente, a extração do cordão de solda na detecção automática de defeitos permite restringir o local da busca por descontinuidades no cordão. Essa restrição acelera a busca e evita que objetos que não estejam no interior da região desejada sejam detectados como defeitos (evita falsos positivos fora do cordão). Algumas abordagens se baseiam na segmentação da região do tubo como etapa preliminar, outras delimitam diretamente a região do cordão de solda.

Park et al. [22] apresentam uma técnica para segmentação da região do cordão de solda em imagens radiográficas do tipo PDVD pela extração dos perfis do tubo. Na etapa de pré-processamento é feita a remoção dos ruídos através da aplicação do filtro da média, o que reduz a intensidade dos picos no momento da extração dos perfis. Em seguida os perfis verticais são extraídos de três posições diferentes do tubo representando o fundo, o cordão de

solda e a espessura do tubo. O perfil horizontal é extraído linha a linha, o qual em conjunto com os outros perfis estabelece os limites do tubo. A junção entre os perfis horizontais e verticais permite identificar a posição do cordão de solda. Os testes desenvolvidos pelos autores indicam que o método proposto pode ser aplicado às imagens radiográficas com baixo contraste, borrada e com ruídos.

Kroetz [4] e Kroetz et al. [12] realizam a detecção do cordão de solda em duas etapas. Primeiramente, a região do tubo é detectada e em seguida faz-se a busca pela região do cordão de solda dentro dos limites tubo. Em [4] e [12], a detecção do tubo utiliza técnicas bioinspiradas como algoritmos genéticos (AG) e otimização por enxame de partículas (PSO do inglês *Particle Swarm Optimization*) na produção de soluções candidatas. Estas soluções devem ser tais que maximizem a similaridade de um determinado recorte na imagem alvo em comparação com uma imagem modelo do tubo. O modelo do tubo é construído a partir de seis parâmetros que descrevem a posição do tubo na imagem como posições horizontal e vertical, largura, extensão, espessura e orientação. A partir desses atributos os algoritmos bioinspirados encontram a posição do tubo na imagem.

O trabalho desenvolvido em [12] utiliza imagens do tipo PDVD e obtém uma taxa de acerto no reconhecimento total do tubo de 75% em 8 imagens. Em [4] o conjunto de imagens testadas é ampliado para 117 imagens do tipo PDVD e a abordagem apresenta como resultado final 20,51% de acerto na detecção do tubo. Em [4] tal resultado influi diretamente na detecção do cordão de solda, pois se a imagem do tubo não é detectada corretamente, não é possível buscar a região do cordão de solda.

Suyama e colaboradores [13] e [14] apresentam uma técnica para auxiliar na localização do cordão de solda em imagens radiográficas do tipo PDVD, buscando a região central do cordão de solda. Os autores se baseiam no fato de que a região central do cordão de solda em imagens radiográficas do tipo PDVD tem formato elíptico. Inicialmente, são aplicadas técnicas de processamento de imagens para segmentar regiões (chamadas regiões candidatas) nas quais é possível inscrever uma elipse. Em seguida é usada uma técnica baseada em PSO para buscar a região que permite inscrever a elipse com a maior área possível. A região encontrada é definida como a região central do cordão de solda. Ambos os trabalhos usam nove imagens para a validação da técnica. Em [13], o percentual médio de acerto na classificação das regiões candidatas é de 77%. Em [14], no qual se utiliza a técnica de PSO também para balancear os pesos dos parâmetros de busca, o percentual médio de acerto na detecção do cordão é de 88%.

Embora as abordagens descritas em [13] e [14] busquem regiões elípticas na imagem que se assemelham à forma do cordão de solda, estas não são capazes de detectar apenas o cordão, mas apenas auxiliar na busca do cordão de solda. Os trabalhos apresentados em [13] e [14] não fazem a detecção prévia da região do tubo.

3 Metodologia Proposta

Neste trabalho propõe-se uma nova técnica de segmentação automática do tubo cujas etapas do processo são ilustradas na figura 2. O algoritmo foi desenvolvido no *software Matlab*® versão R.14a e o *Image Processing Toolbox* em um computador com processador Core i7 e 8 Gb de memória.

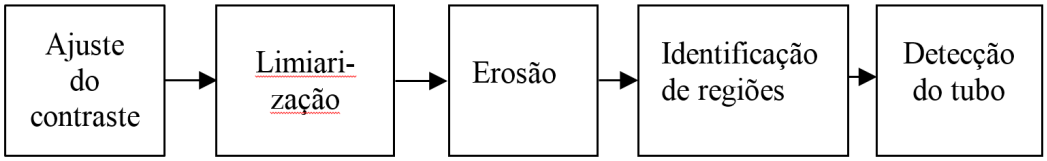


Figura 2. Processo de detecção do tubo.

A figura 3(a) mostra uma imagem radiográfica do tipo PDVD contendo o tubo a ser segmentado. Observa-se que a região de fundo é constituída por uma parte escura contornada por um retângulo branco. O processo de detecção do tubo é realizado pelas etapas descritas a seguir.

Ajuste do contraste. Nesta etapa é aplicada a transformação logarítmica definida na equação (1). Essa transformação ajusta um intervalo expandindo os valores de intensidade mais baixos em um intervalo maior de níveis de saída com o objetivo de destacar a área do tubo do fundo da imagem. A figura 3(c) apresenta a região do tubo de forma mais clara em relação à imagem original.

Limiarização. Após o ajuste do contraste, a imagem do tubo se torna mais nítida permitindo aplicar uma limiarização que converte a imagem para apenas dois níveis de cinza. A técnica de limiarização por Otsu [18] procura separar a imagem em duas classes distintas (Figura 3(e)). Observa-se que a limiarização não é suficiente para segmentar somente a área do tubo. Em algumas das imagens radiográficas, devido ao posicionamento do filme radiográfico, a região branca que deve pertencer ao fundo se conecta com a região do tubo como observado na parte inferior da figura 3(b). Para minimizar os efeitos da limiarização por Otsu a imagem foi normalizada através da equação 3 que restringe faixas de intensidade para os valores desejados. Neste caso, às faixas com pequeno valor de intensidade são atribuídas o valor zero (preto), enquanto que as demais faixas próximas ao maior valor de intensidade passam a ter o valor máximo de intensidade da imagem (*max*). O valor *max* varia conforma a escala dinâmica da imagem podendo estar no intervalo $[0, 2^8 - 1]$ ou no intervalo $[0, 2^{16} - 1]$ (Figura 3(d)).

$$g(x,y) = \begin{cases} 0 & \text{se } f(x,y) \geq 0.98 * max \\ max & \text{se } 0.2 * max < f(x,y) < 0.98 * max \end{cases} \quad (3)$$

Erosão. Para evitar a conectividade entre as regiões de fundo e do tubo e para reduzir pequenos objetos que possam interferir no processo de etiquetagem foi aplicada a erosão com um elemento estruturante (EE) retangular com tamanho 15x15, que além de remover as linhas finas reduziu a conexão entre regiões. Este operador morfológico permite que partes da imagem sejam reduzidas ou excluídas, como é o caso das linhas que aparecem na figura 3(e).

Identificação de regiões. Na imagem resultante, mesmo após o recorte, podem permanecer elementos indesejáveis que devem ser excluídos da segmentação final. Para tal aplicou-se um algoritmo de etiquetagem (*labelling*) que atribui um rótulo a cada região identificada como objeto na etapa anterior. A etiquetagem permite saber quantas regiões foram detectadas como objeto bem como a quantidade de pixels que cada região contém. Para a geração desta técnica foi utilizada a função *bwlabel()* do *Image Processing Toolbox* do *Matlab*®.

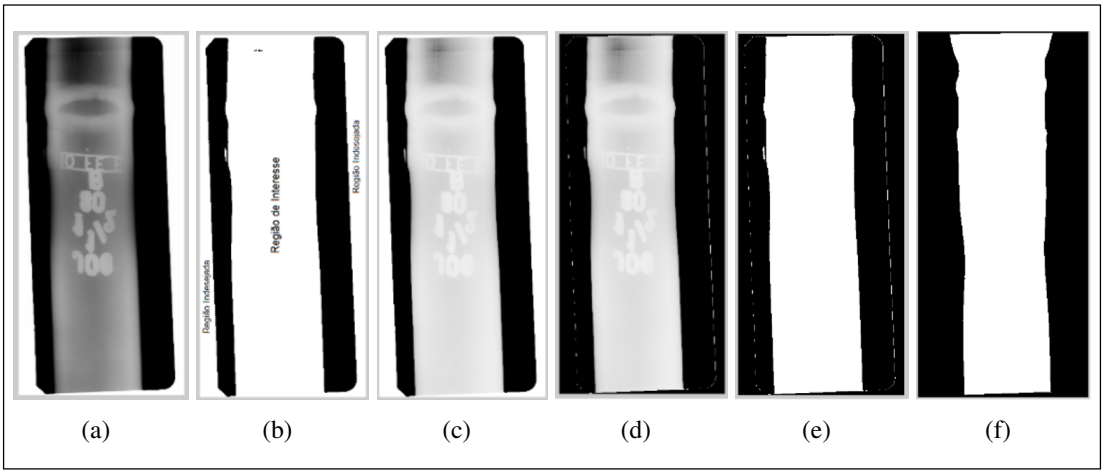


Figura 3. Resultado da aplicação das técnicas: (a) sobre a imagem original, (b) região de interesse, (c) ajuste do contraste, (d) limiarização por faixa dinâmica, (e) limiarização por OTSU e (e) erosão.

A figura 4(b) mostra as regiões reconhecidas em cinza com contorno em branco. Dessa forma, é possível calcular a área de cada objeto detectado na imagem pela sua quantidade de pixels. O objeto com a maior área é designado como sendo a região do tubo, eliminando assim a informação irrelevante (figura 4(b)). O resultado final é a detecção do tubo em relação a imagem original (figura 4(c)).

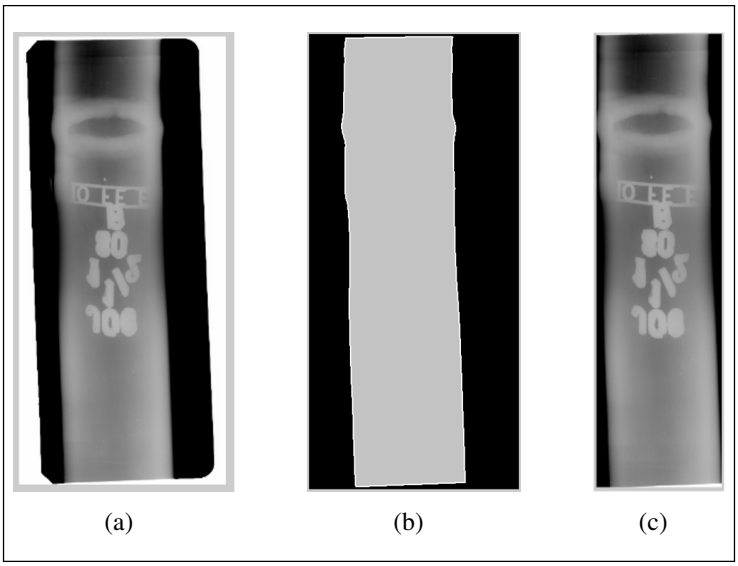


Figura 4. Processo de etiquetagem em a (a) Imagem original, (b) Etiquetagem e (c) Região do tubo detectado.

4 Resultados e Discussões

O método proposto neste trabalho foi aplicado sobre um conjunto de 117 imagens radiográficas digitais formato "*TIFF*", das quais 16 foram obtidas pelo sistema Dürr, 34 pelo sistema GE e 65 pelo sistema Kodak. As imagens Dürr possuem intensidade 16 bits e as demais 8 bits. O sistema proposto obteve neste conjunto de imagens uma precisão de 99,14% na identificação e extração do tubo. Todas as imagens do tudo estão dispostas verticalmente.

A qualidade das imagens utilizadas na base de dados varia de acordo com um conjunto de variáveis, entre elas o tipo de equipamento, tempo de exposição, filtros aplicados, tipo do filme, tipo do equipamento de digitalização, entre outros elementos [23]. Esses parâmetros influenciam não apenas no contraste da imagem como também na quantidade de ruído e na posição do tubo na imagem de forma que seria necessário considerar esses aspectos no algoritmo para que a porcentagem de acerto fosse maior.

A avaliação do método proposto levou em consideração o fato de que se o cordão de solda é mantido intacto na detecção do tubo o método atinge o objetivo e o resultado pode ser levado à próxima etapa, que é a detecção do cordão. Neste caso, o resultado é considerado como um acerto. Se a detecção do tubo perder o cordão de solda ou parte dele o resultado é insatisfatório, pois a etapa seguinte do processo não poderá detectar o cordão de solda. Todo o processo avaliativo foi executado através de análise visual envolvendo ambos os sistemas: Kroetz [4] e o método proposto.

A figura 5(a) apresenta a imagem original, as demais imagens são as possíveis falhas na detecção do tubo, as quais foram recortadas manualmente para exemplificar os possíveis erros na detecção. Neste sentido, busca-se manter intacta a região do cordão de solda tanto no sentido vertical quanto horizontal. As figuras 5(b), 5(d), 5(e) e 5(f) mostram que parte do cordão de solda foi removido, enquanto que na figura 5(c) a região do cordão foi extraída completamente.

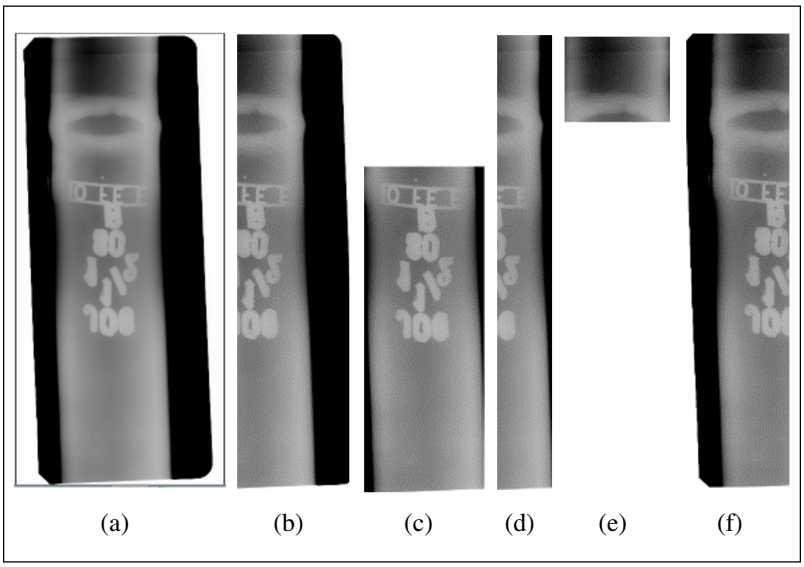


Figura 5. Imagens de referência com erros de detecção: (a) imagem original, (b), (c), (d), (e) e (f) erros de detecção.

Na figura 6 são apresentadas 4 imagens resultantes do processo de detecção do tubo através do sistema de Kroetz [4] relacionadas ao processo de avaliação utilizado. A figura 6(a) mostra uma detecção parcial do tubo com corte do cordão gerando um erro de detecção. Em 6(b) e 6(c) a região detectada é maior que a região do tubo, mas mantém a integridade do cordão de solda. A figura 6(d) apresenta a detecção ideal, ou seja, a região reconhecida equivale à área do tubo mantendo a região do cordão intacta.

Para uma melhor compreensão das diferenças entre os dois sistemas, a Figura 7 mostra o mesmo conjunto de imagens obtidos pelo sistema proposto após aplicação do mesmo sistema de avaliação visual. As imagens em 7(a), 7(b) e 7(c) detectam uma região maior que a área do tubo, mas mantém a região do cordão de solda, em 7(d) tem-se a detecção ideal.

A figura 8 mostra uma falha na detecção do tubo pelo método proposto em relação ao método de Kroetz [4], pois a região do cordão de solda foi eliminada. O método de Kroetz [4] faz a detecção do tubo na mesma imagem mantendo a região do cordão de solda.

A metodologia proposta foi comparada com resultados obtidos pelo sistema de Kroetz

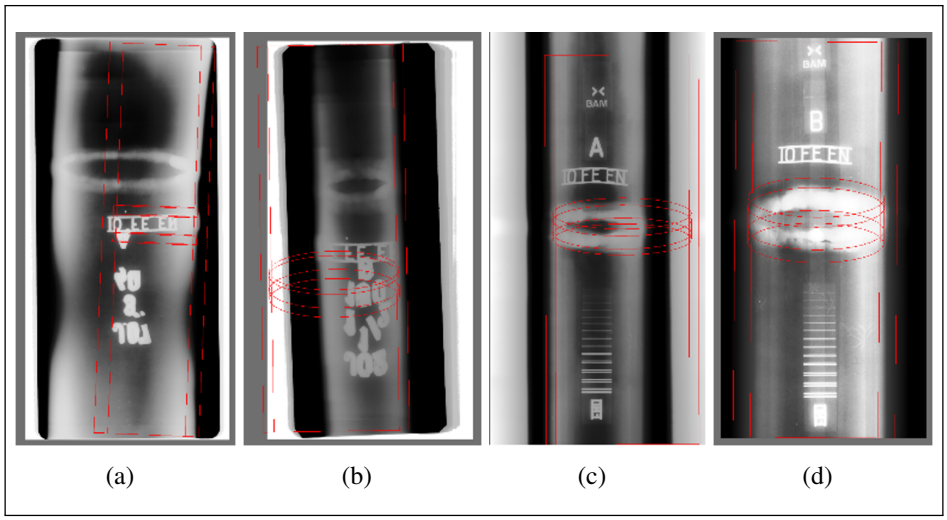


Figura 6. Imagens resultantes do processo visual para identificação da região do tubo pelo método de Kroetz [4]. (a) Erro de detecção, (b) detectado, (c) detectado e (d) detectado.

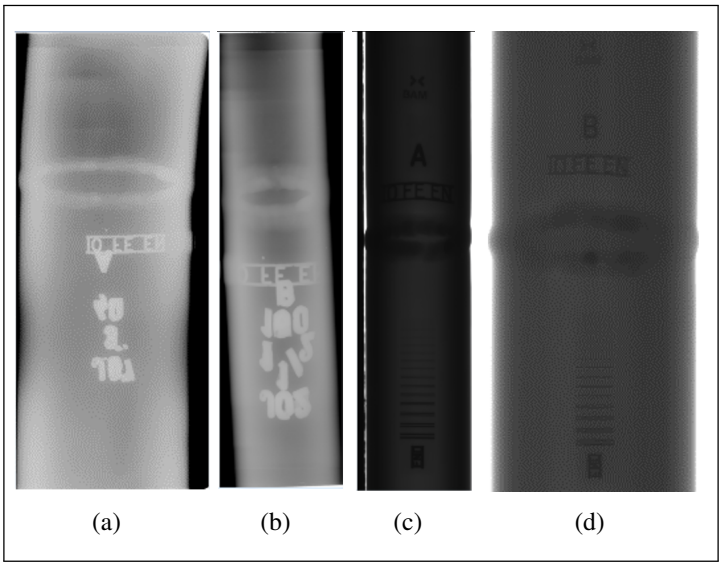


Figura 7. Imagens resultantes do processo visual para identificação da região do tubo pelo método proposto, em (a) detectado, (b) detectado, (c) detectado e (d) detectado.

[4] utilizando o mesmo conjunto de imagens. Neste caso, no geral o sistema de Kroetz [4] obteve uma precisão de 82,05% na detecção do tubo em relação ao conjunto total de imagens radiográficas, contra 99,14% do sistema proposto neste trabalho (Tabela 5). O sistema desenvolvido por Kroetz necessita de interação do usuário na etapa de pré-processamento

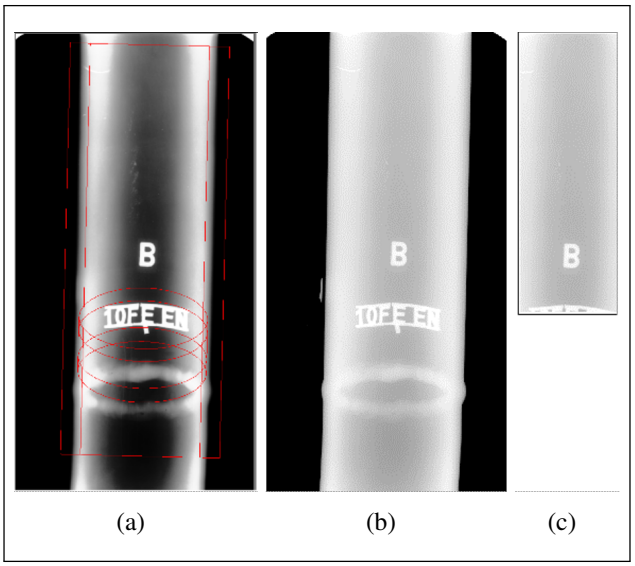


Figura 8. Resultados de detecção do tubo; (a) detecção pelo método Kroetz [4] avaliada como detectado, (b) imagem original para método proposto e (x) falha na detecção pelo método proposto.

e esta etapa depende do conhecimento prévio do usuário na seleção dos ajustes ideais. No teste foi utilizada a seguinte sequência de pré-processamento definida pelo autor [4]: inversão da imagem, equalização do histograma, aplicação do filtro da média com operador 3x3 para posteriormente executar a busca do tubo.

Para uma melhor compreensão dos resultados foram utilizadas 5 tabelas, onde as tabelas 1, 2 e 3 mostram respectivamente a quantidade de acertos para cada classe em relação aos sistemas Dürr, GE e Kodak. A tabela 4 apresenta os dados agrupados em relação à quantidade total de imagens utilizadas.

A tabela 5 apresenta o percentual para cada sistema e também o resultado total em relação ao conjunto de imagens utilizadas. Nesta tabela observa-se que o sistema proposto teve uma melhora significativa em relação ao método de Kroetz [4]. As tabelas mostram que o sistema proposto melhorou a detecção e extração da região do tubo, promovendo um avanço em relação à abordagem anterior.

5 Conclusões

A detecção do cordão de solda é uma etapa essencial para a realização de um processo automático de identificação de defeitos em imagens radiográficas de juntas soldadas de tu-

Tabela 1. Classificação das imagens do sistema Dürr

	Kroetz[4]	Sistema Proposto
Erro	4	0
Detectado	12	16
total	16	16

Tabela 2. Classificação das imagens do sistema GE

	Kroetz[4]	Sistema Proposto
Erro	4	0
Detectado	31	35
total	35	35

Tabela 3. Classificação das imagens do sistema Kodak

	Kroetz[4]	Sistema Proposto
Erro	13	1
Detectado	53	65
total	66	66

Tabela 4. Classificação de todas as imagens de teste

	Kroetz[4]	Sistema Proposto
Erro	21	1
Detectado	96	116
total	117	117

Tabela 5. Taxa de detecção do tubo

	Kroetz[4]	Sistema Proposto
Dürr	75%	100%
GE	88,57%	100%
Kodak	80,30%	98,48%
total	82,05%	99,14%

bulações. Com o objetivo de reduzir o espaço de busca na detecção do cordão é desejável identificar previamente a posição do tubo. O método apresentado neste artigo utiliza uma abordagem simples e rápida para realizar a segmentação do tubo em imagens radiográficas de cordões de solda do tipo PDVD e pode ser utilizado como uma etapa anterior à detecção do cordão de solda.

O resultado alcançado foi de 99,14% no conjunto das imagens testadas (117 imagens) em função das amplas variações de contraste, da quantidade de ruído e do posicionamento do tubo, causadas por diferentes parâmetros no momento da aquisição e sistemas Dürr, GE e Kodak. Houve um avanço do método quando confrontado com o método proposto anteriormente por Kroetz [4] para o mesmo conjunto de imagens utilizadas como teste.

Em trabalhos futuros pretende-se incluir no algoritmo modificações que possam contemplar variações nas imagens decorrentes de diferentes ajustes de parâmetros na aquisição e incluir a etapa de segmentação do cordão de solda.

Contribuição dos autores:

- Marlon de Oliveira Vaz: concepção, planejamento, análise ou interpretação dos dados.

- Tania Mezzadri Centeno: revisão crítica de importante conteúdo intelectual e aprovação final da versão a ser publicada.

- Myrian Regattieri Delgado: revisão crítica de importante conteúdo intelectual.

Referências

[1] FüCSöK, F.; MULLER, C.; SCHARMACH, M. Human factors: the nde reliability of routine radiographic film evaluation. In: . c2000. v. 3.

[2] LIAO, T. W.; NI, J. An automated radiographic ndt system for weld inspection: part i - weld extraction. *Ndt & E International*, v. 29, n. 3, p. 157–162, 1996.

[3] FELISBERTO, M. K. *Técnicas automáticas para detecção de cordões de solda e defeitos de soldagem em imagens radiográficas industriais*. 2007. Tese (Doutorado em Engenharia Elétrica e Informática Industrial) - UTFPR, 2007.

[4] KROETZ, M. G. *Sistema de apoio na inspeção radiográfica computadorizada de juntas soldadas de tubulações de petróleo*. 2012. Dissertação (Mestrado em Engenharia Elétrica e Informática Industrial) - UTFPR, 2012.

[5] ZAPATA, J.; VILAR, R.; RUIZ, R. Performance evaluation of an automatic inspection system of weld defects in radiographic images based on neuro-classifiers. *Expert Systems with Applications*, v. 38, n. 7, p. 8812–8824, 2011.

[6] XIAOMENG, W. Detection of weld line defect for oilgas pipeline based on x-rays image processing. In: . c2009. p. 273–275.

[7] HALIM, S. A.; HADI, N. A.; IBRAHIM, A.; MANURUNG, Y. H. The geometrical feature of weld defect in assessing digital radiographic image. In: . c2011. p. 189–193.

[8] LIAO, T. W.; TANG, K. Automated extraction of welds from digitized radiographic images based on mlp neural networks. *Applied Artificial Intelligence*, v. 11, n. 3, p. 197–218, 1997.

[9] LIAO, T.; LI, D.-M.; LI, Y.-M. Detection of welding flaws from radiographic images with fuzzy clustering methods. *Fuzzy sets and Systems*, v. 108, n. 2, p. 145–158, 1999.

[10] LIAO, T. W. Fuzzy reasoning based automatic inspection of radiographic welds: weld recognition. *Journal of Intelligent Manufacturing*, v. 15, n. 1, p. 69–85, 2004.

[11] FELISBERTO, M. K.; LOPES, H. S.; CENTENO, T. M.; DE ARRUDA, L. V. R. An object detection and recognition system for weld bead extraction from digital radiographs. *Computer Vision and Image Understanding*, v. 102, n. 3, p. 238–249, 2006.

[12] KROETZ, M.; CENTENO, T. M.; DELGADO, M. R.; FELISBERTO, M.; LUCAS, L. A.; DORINI, L. B.; FYLYK, V.; VIEIRA, A. Genetic algorithms to automatic weld bead detection in double wall double image digital radiographs. In: . c2012. p. 1–7.

[13] SUYAMA, F. M.; KREFER, A. G.; FARIA, A. R.; CENTENO, T. M. Identificação da região central de cordões de solda em imagens radiograficas de tubulações de petróleo do tipo parede dupla vista dupla. In: . c2013.

[14] SUYAMA, F. M.; KREFER, A. G.; FARIA, A. R.; CENTENO, T. M. Detecting central region in weld beads of dwdi radiographic images using pso. *International Journal of Natural Computing Research (IJNCR)*, v. 5, n. 1, p. 42–56, 2015.

[15] BOILER, A.; CODE, P. V. *Section v: Nondestructive examination*. Subcommittee on Fiber-Reinforced Plastic Pressure Vessels, 2011.

[16] GONZALEZ, R. C.; WOODS, R. E. *Processamento digital de imagens. tradução: Cristina yamagami e leonardo piamonte*. Pearson Prentice Hall, São Paulo, 2010.

[17] SEZGIN, M. et al. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, v. 13, n. 1, p. 146–168, 2004.

[18] OTSU, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, v. 9, n. 1, p. 54–63, 1979.

[19] SHIH, F. Y. *Image processing and pattern recognition: fundamentals and techniques*. John Wiley & Sons, 2010.

[20] NIXON, M. *Feature extraction & image processing*. Academic Press, 2008.

[21] SHAPIRO, L.; HARALICK, R. *Computer and robot vision*. Addison-Wesley, 1992. v. I.

[22] PARK, S.-K.; AHN, Y.-S.; GIL, D.-S. The study on evaluation for digital radiography image of weldmentsthe study on evaluation for digital radiography image of weldments. In: . c2011.

[23] DA SILVA, R. R.; MERY, D. State-of-the-art of weld seam inspection by radiographic testing: part i–image processing. *Materials Evaluation*, v. 65, n. 6, p. 643–647, 2007.

Método Sistêmico com Suporte em GORE para Análise de Conformidade de Requisitos não Funcionais Implementados em Software

Systemic Method Supported by GORE for Analysis of Software-Implemented Non-Functional Requirements Compliance

André Luiz de Castro Leal [11]

Henrique Prado Sousa [22]

Julio Cesar Sampaio do Prado Leite [33]

Data de submissão: 11/06/2015, Data de aceite: 01-03-2016

Resumo: A análise de requisitos não funcionais (RNF) é um desafio e vem sendo explorado na literatura científica há muito tempo. Tal iniciativa deve-se ao fato da existência do problema de se verificar o uso das operacionalizações desse tipo de requisito no software construído. Nesse trabalho apresentamos um método, com técnicas e ferramentas de apoio, que verificam se um software está em conformidade com padrões de RNF estabelecidos em catálogo como alternativa para o problema de verificação de RNF. A estratégia adotada utiliza agentes autônomos para verificação de conformidade de software em relação a operacionalizações de RNF, para isso utiliza uma base de conhecimentos de padrões persistidos em um catálogo. Os resultados, parciais, são indicativos de que a proposta de solução é aplicável. A avaliação da validade dá-se por demonstração de que um método parcialmente automatizado que é eficaz na identificação de conformidades. Um diferencial na proposta é que, de maneira geral, os trabalhos com foco em verificação são fortemente voltados para a visão funcional, enquanto a solução aqui apresentada é inovadora na ligação dos RNF a sua efetiva implantação. Como prova de conceito aplicou-se e customizou-se uma técnica de padrões de RNFs baseados em orientação a objetivos em estudos de caso de exemplos do cotidiano prático de software. Como também a construção de um framework de agentes, que operam sob notações XML, para identificar conformidades de software em relação a um catálogo de RNF.

Palavras-chave: conformidade, requisito não funcional, orientado a objetivos, NFR-framework

1 André Luiz de Castro Leal, D.Sc. - Universidade Federal Rural do Rio de Janeiro – UFRRJ - Seropédica – RJ {andrecastr@gmail.com}

2 Henrique Prado de Sá Sousa, M.Sc. - Universidade Federal Rural do Rio de Janeiro – UFRRJ - Seropédica – RJ. {hsousa@inf.puc-rio.br}

3 Julio César Sampaio do Prado Leite, D.Sc. - Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio - Rio de Janeiro – RJ. {jcspl@inf.puc-rio.br}

Abstract. The analysis of non-functional requirements (NFR) is a challenge and has been explored in the literature for a long time. This initiative is due to the fact of the existence of the problem of verifying the use of the NFR's operationalization in software. In this paper WE present a method, with supporting tools and techniques that check if a software complies with standards of non-functional requirements as described in catalog, as an alternative to the NFR verification problem. The strategy adopted in this thesis uses autonomous agents to check software compliance regarding THE operationalization of nfr, by using a knowledge base of *patterns* persisted in catalog. partial results show that the proposed solution is applicable. The evaluation of the validity is given by the demonstration that a partially automated method is effective in identifying compliance. our approach adds strongly over existing approaches in that it innovates in linking NFR to its effective implementation, while existing approaches focus on a functional view. For demonstration, it was applied and customized it a technique based on *patterns* of NFRs oriented goals on case studies examples of practical everyday software. As well as the construction of a framework for agents that operate in XML notation to identify compliance of software with respect to a catalog of RNF.

Keywords: compliance, non-functional requirements, goal oriented, NFR-framework

1 Introdução

Requisitos de restrição e qualidade em software propõem características em alto nível de abstração e sua avaliação possui grau de julgamento que depende do ponto de vista do interessado [1]. Por exemplo, dizer se é transparente ou não, tomado Transparência como um Requisito não Funcional (RNF) [2], pode não ser adequado, pois irá depender do grau de subjetividade de quem está na expectativa da informação. Além disso, conforme apresentado no Catálogo de Transparência de Software (CTS) (2013), a Transparência possui decomposições em outros RNFs e elos de relação entre eles que podem sugerir um impacto de julgamento ainda mais complexo, uma vez que tais relações podem ser positivas ou negativas [3]. Essas decomposições são tratadas de metas-flexíveis a partir do framework de Chung et al. (2000) [4].

Enquanto um Requisito Funcional (RF) é uma declaração do que o software deve ou não deve fazer, normalmente expressa na forma: se uma dada condição é, em seguida, o software deve responder apropriadamente; um RNF é uma qualidade ou restrição inerente ao comportamento do software diante dessa resposta [5]. A verificação dessas condições torna-se mais efetiva para os RFs, uma vez que dada a condição, a verificação e validação da resposta dada pelo software é menos passível de controvérsia, pois é determinística. Já para RNFs a validação de sua implementação é algo mais suscetível ao olhar de quem analisa. A implementação de uma determinada qualidade pode satisfazer ao julgamento de um usuário, mas pode não estar adequada às expectativas de outro usuário. Além disso, é composta por características que não podem ser implementadas diretamente, em vez disso são satisfeitas em alguns casos pela combinação de diversos RFs [5].

A literatura apresenta que RNFs são difíceis de se expressar e consequentemente de verificar e validar nas diversas fases do software porque: certas restrições estão relacionadas com a solução de desenho, que é desconhecido na fase requisitos [6]; certas restrições, são altamente subjetivas e só podem ser determinadas através de avaliações empíricas complexas [7]; RNFs podem se relacionar a um ou mais RFs e possuem dessa maneira uma característica de ortogonalidade [2]; RNFs tendem a conflitos e contradições, uma vez que são suscetíveis à percepção de quem o avalia [8]; não há regras "universais" e diretrizes para determinar quando RNFs são plenamente atendidos [9].

Analisar a implementação em software de RNF isoladamente a partir de sua natureza semântica ou conceitual é uma tarefa difícil diante das características apresentadas, portanto faz-se necessário refinar esses requisitos até que se tornem elementos mais concretos de julgamento [4]. Fenton e Pfleeger (1997) [1] sugerem que RNFs devam ser subdivididos em um conjunto de características funcionais, para que essas possam ser implementados em software para atender às expectativas do usuário com relação aos RNFs elicitados. Lamsweerde (2001) [10] e Lamsweerde e Letier (2000) [11] sugerem especificações hierárquicas de decomposição de metas produzindo estados a serem alcançados e do comportamento do sistema necessário para atingir esses estados. Lamsweerde (2001) [10] argumenta que de modo considerável o refinamento de RNF é necessário para que o raciocínio automatizado possa ser aplicado. Estabelecer um método que faça uma união de abordagens de definição de RNFs, seus refinamentos em questões e operacionalizações transformados em catálogo, para minimizar o grau de subjetividade do julgamento dos RNFs é um desafio. O termo catálogo de RNF define a especialização em camadas de padrões (*patterns*) que envolvem desde requisitos de alto nível de abstração como qualidades RFs de mais baixo nível ou operacionalizações [4] [12]. O objetivo de sua criação em *patterns* é possibilitar reuso de seu conhecimento. Supakkul et al. (2010) [12] definem diferentes tipos de *patterns* tais como: padrões objetivos, padrões problema, padrões alternativas e padrões seleção que serão apresentados em seções posteriores.

Chung et al. (2000) [4] atuam no desafio quando propõem um framework para representar RNF e suas relações; em [12], que definem padrões de tratam desde o conceitual até o operacional para que RNFs fossem representados e avaliados com relação a seu grau de satisfação; e em [13] sugerem a adaptação da proposta de [12] quando exploram no modelo *Goal-Question-Metrics* (GQO) o tratamento do último nível do *Goal-Question-Metrics* (GQM) como operacionalizações baseadas em práticas da Engenharia de Software (ES).

Dentro do contexto de pesquisas de análise de conformidade de implementação de RNFs em software, alguns pontos de investigação se destacam:

- Proposta de uma arquitetura única: um ponto pouco explorado de pesquisa é tratar *patterns* em uma única arquitetura que possa servir de base para automação da análise de RNFs. Tal abordagem mostra-se importante, uma vez que analisar os desmembramentos dos RNFs e suas relações, bem como as alternativas de operacionalizações podem não ser triviais necessitando assim de uma estrutura que

viabilize a analise de satisfação de RNFs. A proposta é importante para que se dê foco ao reunso de uma arquitetura na análise de diferentes RNFs.

- Estabelecimento de um método: no sentido de se tornar possível a abordagem sugerida para análise de RNF, o objetivo primeiro da pesquisa é estabelecer um método a partir de uma abordagem *top-down* ao especificar RNF desde modelos conceituais até seu detalhamento em direção às operacionalizações, até que possam ser analisadas por agentes autônomos.

- Uso de catálogo de RNFs: O trabalho toma como insumo o catálogo definido em [47] que possui um desdobramento de vários RNFs a partir do requisito de Transparência, tais como: Acessibilidade, Usabilidade, Informativo, Entendimento, Auditabilidade e suas decomposições. Dessa forma, o catálogo proposto em CTS (2013) disponibiliza uma série de RNFs caracterizados conceitual e operacionalmente que podem permitir uma análise da sua implementação em software.

Até esse ponto de elicitação, especificação e refinamento das qualidades, o catálogo é concebido a partir da proposta do NFR *Framework*, tratado no trabalho como RNF *framework* [4]. O *framework* propõe a representação das qualidades e seus elos de contribuição positivos ou negativos, além de seu nível de satisfação (satisfeito a contento). Além disso, o *framework* permite a representação da concepção e raciocínio dos processos em grafos, chamados *Softgoal Interdependency Graph* (SIG) e oferece um catálogo de conhecimento, incluindo técnicas de desenvolvimento. O modelo ainda permite explicitar as operacionalizações que podem tornar os RNFs, tratados no *framework* como *softgoals* (metas-flexíveis), como satisfeitos a contento. Já no ponto de operacionalizações há a adoção do GQO [13]. Assim como o GQM [14], o GQO procura elencar elementos que possam tornar as metas, ou no caso RNFs, susceptíveis a avaliação. O GQO utiliza de uma abordagem do uso de boas práticas de ES em sua camada Operacionalization (O). Dessa forma, a camada da criação de operacionalizações procura tornar elementos plausíveis de análise ou julgamento, como por exemplo, utilizar de alternativas de RFs que deem suporte aos RNFs estabelecidos nas camadas de *Goal* (G).

- Adoção da abordagem de *Goal Oriented Requirements Engineering* (GORE): o trabalho constitui-se a partir de uma abordagem orientada a objetivos GORE [10], uma vez que abordagens tradicionais [15], [16] e [17] focam o desenvolvimento apenas na análise dos processos e dados e não capturam o conhecimento (*rationale*) para o software a ser desenvolvido. Isso em muitos casos não contempla toda a complexidade do domínio do problema onde o software está envolvido [18]. Além disso, GORE foca em atividades que precedem a formulação de requisitos do software. As principais atividades na abordagem de GORE perpassam por elicitação e refinamento de metas, suas análises e a associação de metas a agentes [18], tais agentes trabalham de forma conjunta, apesar de sua autonomia, para atingir as metas definidas. Portanto, GORE é uma abordagem que dá suporte ao método a ser proposto.

- Analisar se o uso de SMA é aplicável ao problema de análise de RNFs pré-concebidos em catálogo: A presente pesquisa aproxima-se da abordagem de SMA, uma vez que trata de objetivos a serem atingidos. Agentes em SMA interagem com intuito

de solucionar demandas de sistema, ou metas (*goals*) definidas. O uso de SMA também propõe um potencial para a autonomia das decisões dos agentes em relação a ações para atingir as metas definidas, contribuição entre vários agentes para atingir uma meta comum, estabelecer bases de conhecimento sobre suas ações. Nesse ponto, a pesquisa propõe o uso de agentes com orientação baseada em intencionalidade, ou agentes intencionais [19]. Os agentes baseados em modelos intencionais possuem melhor capacidade de raciocínio e capacidade de aprendizado de agentes inteligentes a partir de bases de conhecimento ou heurísticas, por exemplo. Essas ações inteligentes são baseadas em conceitos como crença, desejo e intenção *(Belief-Desire-Intention* - BDI) [20] [21].

O presente artigo está estruturado da seguinte forma: na seção 2, está a contextualização geral da pesquisa; na seção 3, é apresentado o método; na seção 4, são apresentadas as arquiteturas utilizadas como solução para camada conceitual e de estruturas XML utilizadas no SMA; na seção 5, é apresentada a arquitetura de implementação do SMA; na seção 6, são apresentadas provas de conceito da aplicação do método em análise de RNF em software; na seção 7, estão descritas as conclusões finais.

2 Contextualização

A primeira versão de agentes na análise de RNFs foi um esforço de análise de viabilidade de agentes monitores que atuam de forma conjunta para captura de traços de execução de um software. Após a captura dos dados esses eram comparados com uma lista de regras estabelecidas no CTS. A versão desenvolvida não contemplava uma sistematização de política de monitoração e sim uma modelagem intencional da arquitetura da política de funcionamento dos agentes, isso percebido após a evolução da pesquisa. Seus resultados foram publicados em [22] e [23].

Foi construído um SMA que efetuava coleta de dados dos traços de execução gerados a partir de agentes do software *Lattesscholar* (LS) [24]. O LS é um sistema multi-agentes (SMA) independente para contagem de citações científicas, que combina os serviços do sítio do *Lattes* [25] e do *Google Scholar* [26]. Sua execução é baseada na pesquisa inicial pelo nome do autor no sítio do e posteriormente uma pesquisa pela relação de artigos no *Google Scholar* com o objetivo de contabilizar citações sobre cada publicação do autor. A vantagem de seu uso está na consulta de citações a partir do título da publicação, o que difere, por exemplo do *Publish and Perish* [27], do *Microsoft Academic Search* [28] e do *Google Scholar Citation* [29], uma vez que suas pesquisas são realizadas a partir do nome do autor podendo acarretar inconsistências de contagem devido a erros causados por homônimos, abreviaturas, nomes muito extensos e sinônimos em nomes muito pequenos.

A partir da identificação do software a ser monitorado foi elaborado um modelo intencional baseado em i* (i-estrela) [30] que sugere agentes de monitoração para analisar traços de execução em um software, no caso aplicados sobre o LS.

Os traços de execução do LS são registros gerados em log a cada iteração de execução do LS a partir de pesquisas por nome do autor na plataforma Lattes e o seu número de citações a partir do *Google Scholar*. Os traços de execução possuem registros de cada iteração do LS a partir do armazenamento das ações dos agentes, como *ReceberUrlPesquisador*, *SolicitarObrasCurriculo*, entre outras. Além disso, os traços contém o número de citações por obra do autor. Para reforçar o explicado anteriormente, o LS é um software independente do sistema de monitoração e dele são utilizados apenas os traços de sua execução.

Os traços de execução são monitorados pelos agentes do SMA e as evidências encontradas são comparadas ao catalogo de transparência para indicar conformidades e não conformidades com o mesmo. Os quatro agentes no SMA possuem ações bem definidas e especializadas. Nessa versão, cada agente trata de forma distinta suas tarefas, recursos, desejos e objetivos. Suas discriminações são apresentadas como:

a) MONITOR: o agente monitora constantemente os traços de execução gerados pelos agentes do LS, com o objetivo de captar os seus registros;

b) CANONIZADOR: tem por objetivo receber os registros dos traços de execução e transformar as informações em um padrão que possa ser consolidado.

c) O CANONIZADOR utiliza endereços de posição nos traços de execução, como também ocorrências delimitadas por algum trecho de texto que evidencie o que está sendo canonizado;

d) CONSOLIDADOR: verifica a partir da estrutura canonizada os registros de conformidades e não conformidades e disponibiliza ao ANALISADOR recursos de conformidades que possam ser verificados de acordo com o CTS;

e) ANALISADOR: tem por finalidade verificar se os registros padronizados e extraídos pelo CANONIZADOR correspondem às exigências dos Grupos, Questões e Alternativas normatizadas para um atributo de transparência. O agente verifica quais são as exigências da regra para que uma alternativa seja positivada.

A segunda versão dos agentes veio para corrigir algumas deficiências da primeira. Essas alterações perpassaram por: alteração do formato de mensagens trocadas na comunicação entre os agentes: a comunicação, anteriormente feita a partir de passagem de parâmetros através de métodos das classes da linguagem Java, passaram a ser feitas por estruturas XML, escolhida por ser uma linguagem universal de troca de mensagens, por exemplo, entre webservices; leitura dos traços de execução: na primeira versão eram lidos traços de execução gerados a partir de formatos ASCii[4] e passaram a ser lidos apenas por conteúdos gerados em estruturas XML; outra importante alteração foi o uso de tags lidas das estruturas de comunicação dos agentes

[4] *American Standard Code for Information Interchange*: http://pt.wikipedia.org/wiki/ASCII

e também dos traços de execução baseados em XML: uma vez utilizadas as tag XML entendemos que seu uso poderia facilitar a criação de dicionários de termos que pudessem auxiliar o processo de construção de conhecimento dos agentes. Tags conhecidas e registradas no domínio dos agentes passaram a compor uma base de conhecimento para futuras análises de arquivos com traços de execução de software variados. Tags desconhecidas também passaram a ficar registradas para que pudessem ser analisadas por agente humano e relacionada às regras de Transparência.

3 Modelo central do método sistêmico para análise de RNF

Um modelo central é exposto para que se dê entendimento da completude e complexidade do trabalho, além de deixar explícitas as contribuições a partir do detalhamento das tarefas principais expostas no modelo central. A partir daí as estruturas são apresentadas em blocos para facilitar a leitura e compreensão. A Figura 1 apresenta a partir de *Structured Analysis and Design Technique* (SADT)[5] [17] o modelo central do método sistêmico para análise de RNF. Sua operacionalização é dada a partir de cinco atividades: (A1) Criar SIG (proposto a partir de [4]; (A2) Definir *patterns* (proposto a partir de [12] e [13]; (A3) Configurar XML; (A4) Configurar software (no caso software a ser analisado, cujo papel de configuração fica sob responsabilidade de seu proprietário); (A5) Operacionalizar agentes (propriamente a execução da análise de artefato ou traços de execução do software, os artefatos e traço de execução são tratados simplesmente como artefatos, exceto nos estudos de caso, uma vez que os traços de execução geralmente estão armazenados em arquivos).

O método é apresentado a partir de modelos representados por SADT para deixar explícitas as atividades de sua operacionalização. Foram considerados grandes grupos de atividades na construção do SADT (visão macro) e atividades de menor granularidade foram tratadas em quadros como passos no detalhamento de uma atividade de maior nível. Nos SADT foram incorporadas linhas tracejadas verticais apenas para delimitar atividades e foram inseridos na parte inferior (rodapé) a origem da fundamentação teórica utilizada, nesse caso é indicado a referência bibliográfica do autor que propotos o método ou técnica utilizada em cada passo. A construção do método proposto é portanto uma conjunção de métodos e técnicas propostas por autores da literatura científica da área de ES no que diz respeito a primeira parte da modelagem conceitual dos RNFs, ou seja o que se refere às atividades A1 e A2, a partir de então tais abordagens são associadas e operacionalizadas por outras técnicas inseridas nesse trabalho.

5 Na notação SADT: caixas representam atividades, setas a esquerda representam dados de entrada, setas para baixo representam controles, setas para cima representam mecanismos, e setas para direta representa saídas das atividades.

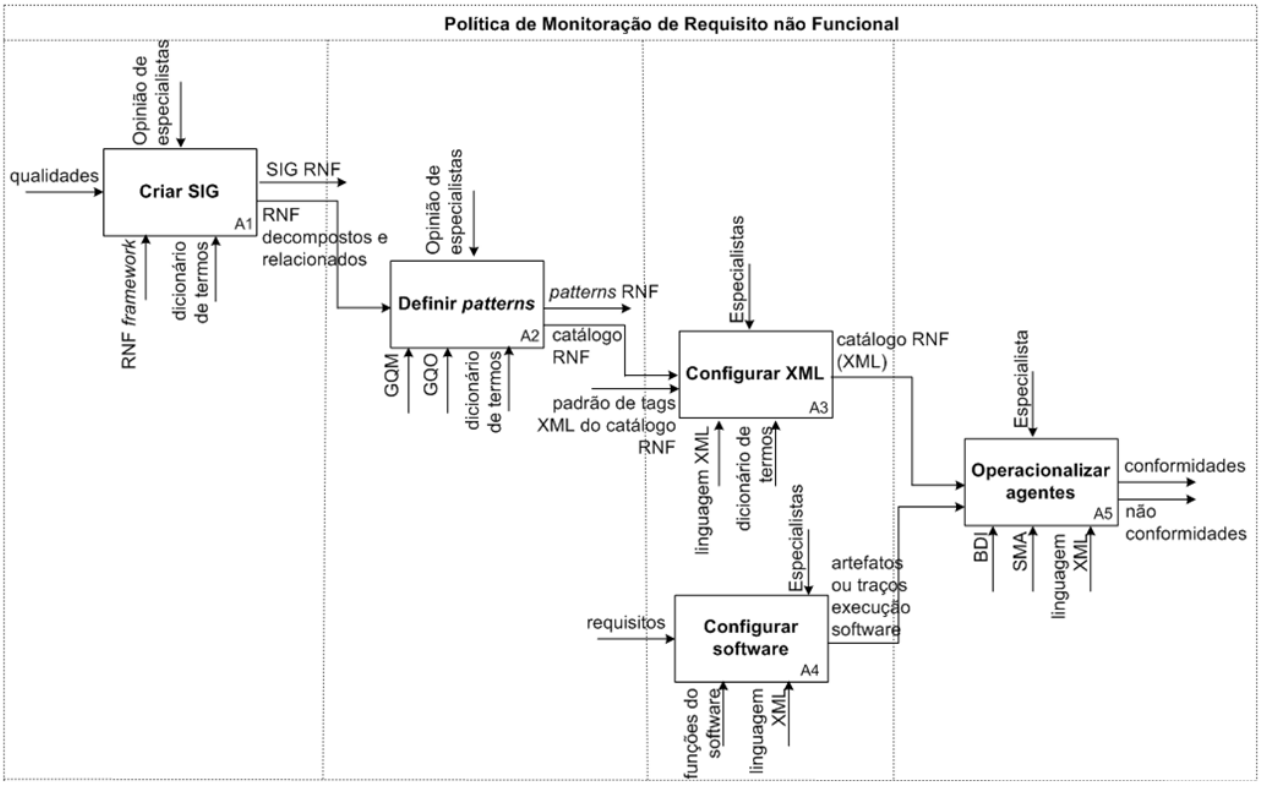


Figura 1. Visão Macro do Método.

1. Criar SIG (A1): consiste em definir de forma sistemática a decomposição de uma qualidade em outras qualidades que a caracterizem, sua hierarquização (com uma das qualidades colocada em evidência, pois trata-se da qualidade que é verificada a partir do modelo) e seus elos de relacionamento, sejam eles de contribuição positivos ou negativos. Além do nível de satisfação (como satisfeito a contento [31], devido as características de subjetividade do atributo). Tal estrutura é baseada na proposta do NFR *framework* de [4]. Esses fatores auxiliam na criação de uma semântica rica para entendimento das qualidades e suas relações e o seu resultado são: RNFs decompostos e relacionados; e SIG do RNF. Em [4] é estabelecido que tais qualidades passam a ser metas a serem cumpridas, ou metas flexíveis. Na ER, o uso da palavra *goal* significa que o método é *goal oriented*, ou seja, trabalha na captura dos requisitos do sistema priorizando a elicitação do porquê (*why*) antes de trabalhar e definir, através dos modelos, o que (*what*) o software deverá fazer [32]. A importância da orientação à meta para o método está no fato de que as qualidades organizadas passam a ser *softgoals* a atingir em um processo de avaliação. Esses *softgoals* devem ser operacionalizados através de procedimentos funcionais que contribuam para sua satisfação, usado com a mesma ideia de *satisfice* (bom o suficiente) [33] para solucionar um problema, e essa satisfação (*satisfice*) é propagada a partir das relações de contribuições entre os demais *softgoals* estabelecidos no modelo SIG. Essa atividade é composta por três atividades: a) Elicitar qualidades: sugere a identificação de um RNF principal e suas decomposições; b) Relacionar qualidades: objetiva relacionar os RNFs identificados/decompostos ao *softgoal* principal; c) Verificar estrutura: sugere com que

as atividades anteriores sejam verificadas. As ponderações ocorridas devem ser verificadas pela atividade onde foi detectada a incoerência.

2. Definir *Patterns* (A2): a atividade tem por objetivo estabelecer maior conhecimento a respeito do *softgoal* objeto da análise a partir de outros *softgoals*, de elos de contribuição, de grupos (categorias), questões e alternativas (operacionalizações/procedimentos funcionais), conforme sugerido no trabalho de [13]. Esses elementos em conjunto permitem a criação de uma estrutura baseada em uma visão *top-down* onde o *softgoal* passa a ser o elemento principal e em uma camada mais baixa estão localizadas as operacionalizações. A importância da aplicação do método está no refinamento a partir do *softgoal* principal e suas decomposições até as operacionalizações que possam satisfazê-los a contento [31]. A partir desses desdobramentos há a criação de um catálogo do *softgoal* a partir de uma coleção de padrões denominado RNF *patterns* (*Objetive patterns, Alternative patterns, Selection patterns, Problem Patterns* [12] e *Question Patterns* [13]). A atividade é sub-dividida em: a) Descrever questões: a atividade tem por objetivo a identificação, descrição e relacionamento de questões que possam responder, com intuito de avaliação; b) Operacionalizar questões: objetiva identificar e relacionar práticas da ES que possam servir de alternativas de respostas para as questões; c) Verificar estrutura: sugere com que as atividades anteriores sejam verificadas. As ponderações ocorridas devem ser verificadas pela atividade onde foi detectada a incoerência.

3. Configurar XML (A3): a partir das duas atividades anteriores o processo de operacionalização do método necessita que os requisitos estabelecidos, bem como suas operacionalizações, sejam configurados para que possam ser utilizados como insumos para o SMA. Essa configuração padroniza estruturas XML, devidamente sistematizadas, hierarquizadas e relacionadas, para que os agentes possam estabelecer um baseline estruturado com um método e avaliação. O arcabouço do XML é apresentado no APÊNDICE A e nele estão registradas as estruturas para a definição de *patterns* do RNF decomposto e relacionado a partir das atividades A1 e A2, dessa forma é estabelecido um padrão para tratamento de conformidades e não conformidades entre um software a ser analisado e as regras definidas no catálogo. O uso do arcabouço XML tem como objetivo possibilitar o reuso do conhecimento sobre RNF [45]. As estruturas em XML dos padrões são mantidas, independente do software analisado, e o ponto de variabilidade encontra-se na configuração do arcabouço XML com a customização dos conteúdos das tags dos padrões dependendo do domínio de aplicação do catálogo. Por exemplo, para o RNF de Transparência [2] há uma customização que envolveria o Universo de Informação (UdI) [34] desse domínio, enquanto que para o RNF de Segurança seria necessária a análise de metas e operacionalizações de seu UdI. A atividade é composta por atividades que tem por objetivo principal estruturar o catálogo do RNF em um padrão XML: a) Configurar *objetive patterns*: objetiva configurar o RNF principal, suas decomposições e relacionamentos (padrões seleção); b) Configurar question *patterns*: objetiva configurar os padrões questões já relacionados às decomposições dos *objetive patterns*; c) Configurar *alternative patterns*: objetiva configurar as operacionalizações (alternativas/respostas) para as questões identificadas para cada decomposição; d) Configurar variáveis e sinônimos: nessa atividade há como

foco a identificação de variáveis e seus sinônimos que servirão de marcadores para configuração do artefato, ou quando necessário, também dos artefatos do software que não são gerados sobre um padrão pré-determinado. Quando há padrão, a configuração pode ser feita apenas no catálogo XML. A atividade consiste, primeiramente na identificação de variáveis e seus sinônimos para cada operacionalização e em um segundo momento na sua configuração em XML; e) Verificar estrutura: consiste em revisar o catálogo configurado para identificar sua consistência, as observações de avaliação devem ser revistas pelos passos anteriores.

4. Configurar software (A4): a importância da atividade consiste em configurar o artefato, ou quando necessário (traços sem padrão pré-determinado), os traços de execução do software que são analisados. Os artefatos devem estar com marcações compatíveis com o padrão estabelecido no XML de configuração do catálogo dos *patterns* de RNF. A atividade é responsabilidade do produtor do software e está representada na regra do catálogo de forma a deixar explícita a necessidade de sua execução enquanto atividade.

5. Operacionalizar agentes (A5): o objetivo da atividade é prover um mecanismo automático de operacionalização da politica de análise a partir de agentes baseados em uma abordagem de SMA. Os agentes são importantes uma vez que trabalham fundamentalmente como sistemas de software que representam os atores em um determinado contexto, são autônomos, pró-ativos e sociais (colaboram entre si, se comunicam e interagem), possuem capacidade de aprendizado e de adaptação [35]. Essas características são importantes em um contexto de análise uma vez que a complexidade dos elementos envolvidos é peculiar. Os agentes trabalham com o contexto do catálogo do RNF parametrizado, conforme sugerido em A3, desde a caracterização e o estabelecimento das relações iniciais dos RNF até os procedimentos funcionais para satisfazê-los a contendo [31]. A atividade é sub-dividida em: a) Capturar artefato: objetiva a captura o artefato do software a ser analisado; b) Analisar: tem por finalidade analisar os artefato a fim de disponibilizá-lo para avaliação; c) Apresentar Resultados: consiste em apresentar os resultados de conformidades, a partir da comparação do artefato, comparando-o a uma estrutura pré-definidas no catálogo XML de RNF. O detalhamento de cada atividade do método pode ser encontrado em [36].

4 Arquiteturas Utilizadas

4.1 Modelo conceitual dos *patterns*

O modelo conceitual dos *patterns* de RNF elaborado nessa seção surge a partir da proposta dos autores: - [12] no que se refere a camadas de "Padrões Objetivos" (*Objective Patterns*), subdividido em *patterns* Identificação e *patterns* Decomposição. O primeiro para identificar o RNF objeto da análise e o segundo elemento suas decomposições. Além dos Padrões Seleção (*Selection Patterns*); - o proposto em CTS (2013) e [13], no que se refere à criação de grupos (categorias) de questões, questões, alternativas (respostas às questões) baseado na proposta do GQO; - variáveis essenciais

e seus sinônimos, adaptado de [37], para que possam servir de marcadores para operações dos agentes do SMA na comparação de artefatos do software analisado com o padrão estabelecido em catálogo;

A Figura 2 apresenta o relacionamento entre os elementos do *pattern* (identificação, decomposições, categorias das questões, quetões e alternativas) e também o elemento variáveis essenciais, onde estão os marcadores que serão objetos de comparação. A figura apresenta sua estrutura e também as cardinalidades de relações.

É importante perceber na estrutura que o *pattern* Identificação é o elemento principal de onde todos os outros passam a ser estruturados. Nele estará o *softgoal* principal pelo qual há a intenção de avaliar seu grau de compatibilidade com as regras configuradas a partir de *patterns*. Esse *pattern* possui decomposições para obter-se maior detalhamento sobre si e esse detalhamento é representado pelo *patterns* Decomposições. Uma decomposição pode ainda ser desmembrada em várias outras decomposições.

Essas decomposições são categorizadas para que alternativas das questões possam ser identificadas e relacionadas a essa categorias para que possam servir para avaliação dos *softgoals*. As questões por sua vez devem possuir alternativas (baseadas em boas práticas de ES, conforme tratado em seções anteriores desse trabalho) para que possam servir como opções de respostas para avaliação dos *softgoals*. As alternativas de respostas deverão ser pensadas com o objetivo de reuso entre as diversas questões do catálogo do RNF, uma vez que os padrões são estabelecidos a partir de uma estrutura que permite seu reuso, onde o ponto de variabilidade encontra-se nos conteúdos preenchidos em cada padrão que dependerâ do domínio da aplicação.

As variáveis essenciais por sua vez devem representar marcadores passíveis de comparação com os artefatos do software. Dessa forma, a comparação entre a heurística e o artefato, bem como sua relação com aos *patterns* alternativas possibilitam a avaliação desses artefato com todas as regras definidas no método.

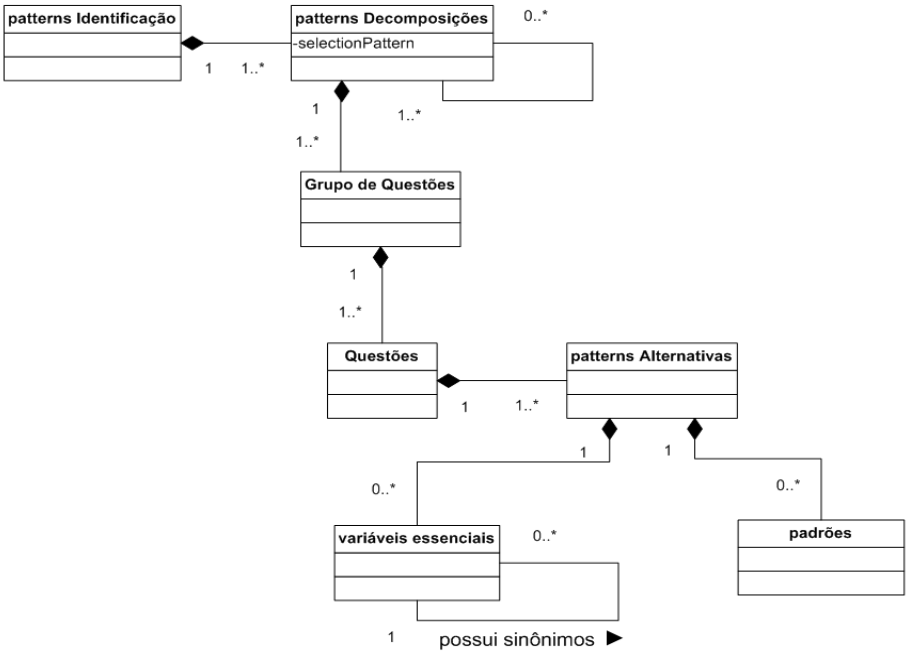


Figura 2. Modelo conceitual dos *patterns* de RNF.

4.2 XML catálogo

Após a criação do modelo conceitual foi possível iniciar a criação de uma estrutura XML do catálogo de *patterns* de RNF, uma contribuição desse trabalho. Sua estrutura é baseada em tags XML criados a partir da estrutura principal do modelo conceitual e serve como regra, uma vez que toda sua arquitetura de decomposição de elementos, relacionamentos, elos de contribuição, conteúdos estão definidas na estrutura.

A estruturas criadas podem ser vista com suas descrições apresentadas na Tabela 1. Na Tabela 1 as tags tags finais são utilizadas conforme padrão XML iniciadas por "/" como por exemplo: </catalogo>, </versão>, </patternObjetivos>, mas não foram apresentadas na tabela.

Tabela 1. Tags presentes no catálogo de RNF.

Identificador	Descrição
<catalogo tipo="" objetivo="">	Identificador inicial da estrutura do catálogo de RNF. Utiliza dois parâmetros TIPO e OBJETIVO. O primeiro indica o tipo de catálogo, caso um catálogo de RNF e o segundo seu objetivo, utilizado nesse trabalho como análise.
<versao>	Versão do catálogo. Deve ser indicada, pois, a análise pode ser baseada em versões diferentes do catálogo.
<patternObjetivos>	Identificador inicial dos *patterns* objetivos.
<patternIdentificacao>	Identificador inicial do RNF principal objeto de análise.

<*patternDecomposicaoNivel_1*>	Identificador inicial da seção quer irá apresentar as decomposições do *patternIdentificacao* em outros RNFs.
<*idDecomposicaoNivel_1*>	Identificador inicial do nome do RNF da decomposição de primeiro nível.
<*patternSelecao*>	Identificador inicial do *patternSelecao* que irá conter elos de ligação do tipo *Some* (+-), *Help*, *Make*, *Hurt*, *Break* que irá indicar o elo do tipo de contribuição da decomposição para o *patternIdentificacao*. Pode estar presente em diferentes seções do XML catálogo. Para o *patternAlternativas* a tag deve ser apenas preenchida com RESPONDE, que indica que a alternativa responde à questão na qual está relacionada.
<*patternDecomposicaoNivel_2*>	Identificador inicial da seção quer irá apresentar as decomposições do *patternDecomposicaoNivel_1* em outros RNFs
<*idDecomposicaoNivel_2*>	Identificador inicial do nome do RNF da decomposição de segundo nível.
<*patternGrupo*>	Identificador inicial das representações das questões e alternativas detalhadas a partir da proposta de GQO.
<*idGrupo*>	Identificador inicial do nome do grupo (agrupador) da questão.
<*tituloGrupo*>	Identificador inicial da descrição do grupo.
<*patternQuestoes*>	Identificador inicial da seção que irá detalhar as questões e suas alternativas.
<*idQuestao*>	Identificador inicial da descrição da questão.
<*tituloQuestao*>	Identificador inicial do título da questão ou descrição da questão.
<*patternAlternativas*>	Identificador inicial da seção que irá detalhar alternativas de respostas para às questões.
<*idAlternativa*>	Identificador inicial do identificador (código ou sigla) da alternativa.
<*tituloAlternativa*>	Identificador inicial do detalhe do título da alternativa.
<*variaveisEssenciais*>	Identificador inicial da seção que irá detalhar as variáveis essenciais e seus sinônimos.

<*idVariaveisEssenciais*>	Identificador inicial da descrição do símbolo (palavra, termo ou frase sucinta) que irá representar uma variável essencial vinculada a uma alternativa.
<*nocaoVariaveisEssenciais*>	Identificador inicial que descreve o significado da variável essencial.
<*sinonimosVariaveisEssenciais*>	Identificador inicial da seção quer irá detalhar os sinônimos das variáveis essenciais.
<*idSinonimosVariaveisEssenciais*>	Identificador inicial da descrição do símbolo (palavra, termo ou frase sucinta) que irá representar um sinônimo vinculado a uma variável essencial.
<*padrão*>	Identificador inicial da descrição de padrões.
<*idpadrao*>	Indicador dos padrões utilizados para representar as alternativas. Os padrões podem ser utilizados com tags simples do tipo: <idpadrao> </idpadrao>, nesse caso o SMA irá efetuar a pesquisa pelo conteúdo indicado entre as tags de forma isolada; e pode utilizar padrões compostos do tipo: <idpadrao1> </idpadrão1>, <idpadrao2> </idpadrão2>, ... <idpadraoN> </idpadrãoN>, o que indica que o SMA irá fazer uma pesquisa e comparação de toda a estrutura do padrão, que deve ser informado de forma seguida, um embaixo do outro, com os conteúdos dispostos entre tags. Para efeito de estudo, está implementado no SMA o uso de % para os conteúdos dos padrões, o que significa que ao encontrar um conteúdo do tipo %, o SMA irá interpretá-lo como qualquer cadeia de string. O uso de mais de um carácter especial % é permitido.

A organização hierárquica do catálogo de RNF pode ser vista conforme apresentado no código XML. O trecho de código é referente à estrutura do metadado das tags explicadas na tabela acima.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<catalogo tipo="" objetivo="">
    <versao>1</versao>
    <patternObjetivos>
        <patternIdentificacao></patternIdentificacao>

        <patternDecomposicaoNivel_1>
            <idDecomposicaoNivel_1></idDecomposicaoNivel_1>
            <patternSelecao></patternSelecao>

                <patternGrupo>
                    <idGrupo></idGrupo>
                    <tituloGrupo></tituloGrupo>
                    <patternQuestoes>
                        <idQuestao></idQuestao>
                        <tituloQuestao></tituloQuestao>
                            <patternAlternativas>
                                <idAlternativa></idAlternativa>
                                <tituloAlternativa></tituloAlternativa>
                                <patternSelecao></patternSelecao>
                                <variaveisEssenciais>
                                    <idVariaveisEssenciais></idVariaveisEssenciais
                                    <nocaoVariaveisEssenciais>  <nocaoVariaveisEss
                                    <sinonimosVariaveisEssenciais>
                                        <idSinonimosVariaveisEssenciais></idSinonimosVariaveisEssenciais
                                        <idSinonimosVariaveisEssenciais></idSinonimosVariaveisEssenciais
                                    </sinonimosVariaveisEssenciais>
                                    <padrao>
                                        <idPadrao1></idPadrao1>
                                        <idPadrao2></idPadrao2>
                                    </padrao>
                                </variaveisEssenciais>
                            </patternAlternativas>
```

Indica o tipo de catálogo

Indica o objetivo do seu uso

Nesse ponto sempre utilizado como RESPONDE, que indica que a alternativa é uma resposta que responde à questão relacionada

4.3 Uso do léxico na definição de variáveis essenciais

A estrutura XML nessa seção apresenta o metadado e o conteúdo das tags referentes às variáveis essenciais e seus sinônimos definidos no XML catálogo. As três tags <*idVariaveisEssenciais*>, <*nocaoVariaveisEssenciais*> e <*sinonimosVariaveisEssenciais*> tiveram como fundamento o que é utilizado em Léxico Ampliado da Linguagem (LAL) [34] para definição de símbolos, seu significado (noção) e sinônimos. O trecho de código é um exemplo dos conteúdos da variável essencial "Data da Coleta" com sua noção e os sinônimos.

```
<variaveisEssenciais>
    <idVariaveisEssenciais>Data da Coleta</idVariaveisEssenciais>
    <nocaoVariaveisEssenciais>Data em que o traço de execução foi coletado
            <nocaoVariaveisEssenciais>
        <sinonimosVariaveisEssenciais>
            <idSinonimosVariaveisEssenciais>Coletado em</idSinonimosVariaveisEssenciais>
            <idSinonimosVariaveisEssenciais>Gravado em</idSinonimosVariaveisEssenciais>
        </sinonimosVariaveisEssenciais>
</variaveisEssenciais>
```

O processo de elicitação do LAL, ou simplesmente, Léxico, permite a identificação das frases e palavras amparadas pelas estratégias de identificação de símbolos da linguagem do problema, ou Universo de Informações (UdI), e da identificação da semântica de cada símbolo [34]. O Léxico do UdI é composto por entradas, onde cada entrada está associada a um símbolo (palavra ou frase) da linguagem do UdI. Esses símbolos podem possuir sinônimos e são descritos por noções e impactos.

O Léxico permite a captura do significado (símbolos) dos termos a partir da Noção, bem como seus sinônimos, como a maioria dos glossários, e também inclui a conotação dos mesmos. Além disso, o Impacto permite descrever os efeitos do uso/ocorrência do símbolo na aplicação, ou do efeito de algo na aplicação sobre o símbolo [34]. Desta forma podemos compreender o significado dos termos de modo independente e em relação a outros termos. Essa definição de noções e impactos devem ser estruturadas com o objetivo de se obter um vocabulário mínimo e circularidade que prescrevem que as noções e impactos devem ser descritos com a utilização dos símbolos da própria linguagem do problema sem o uso demasiado de símbolos externos ao Léxico.

Para identificar a qual classe o símbolo pertence, em [37] são sugeridas as seguintes heurísticas: se o símbolo pratica uma ação, ele é classificado como sujeito; se é quem sofre a ação, é classificado como objeto; se o símbolo é uma situação em um dado momento, ele é classificado como estado; se representa uma ação, é classificado como verbo. A Tabela 2 apresenta os detalhes das classificações de noção e impacto dos símbolos de acordo com as classes.

Tabela 2. Noção e Impacto em Léxico [34].

	Noção	Impacto
Sujeito	Quem é o sujeito	Quais ações executa.
Verbo	Quem realiza, quando acontece e quais os procedimentos envolvidos.	Quais os reflexos da ação no ambiente (outras ações que devem ocorrer) e quais os novos estados decorrentes.
Objeto	Definir o objeto e identificar outros objetos com os quais se relaciona.	Ações que podem ser aplicadas ao objeto.
Estado	O que significa e quais ações levaram a esse estado	Identificar outros estados e ações que pode ocorrer a partir do estado que se descreve.

A abordagem da especificação da linguagem do domínio utilizada nesse trabalho tem por objetivo utilizar a sua estrutura na definição das variáveis essenciais para o método sistêmico. As variáveis essenciais são utilizadas para permitir que o SMA possa fazer a leitura dos artefatos do software, com o objetivo de encontrá-las como marcadores válidos de conteúdos dos artefatos de software e a posterior comparação com as regras definidas. As buscas nos artefatos são feitas a partir das variáveis essenciais e também por seus sinônimos definidos no XML catálogo.

Diante das definições do UdI, o Léxico se torna um potencial elemento para a geração das variáveis essenciais e seus sinônimos, uma vez que esse levantamento e especificação são feitos em tempo de definição de requisitos. Dessa forma, é sugerido que seja feita a associação dos elementos do Léxico ao catálogo XML de RNF.

4.4 Modelo Intencional dos Agentes

Nessa última versão o SMA (as versões anteriores foram apresentadas na seção 3.1 Contexto Histórico da Pesquisa) é composto por quatro agentes com objetivos e funções distintas. Embora haja essa especialização (cada agente com funções distintas e bem definidas) há um objetivo central que é a avaliação das informações dos artefatos de software em relação a um catálogo pré-definido do RNF. Essa avaliação se dá a partir da identificação de conformidades e não conformidades. As conformidades sugerem que o artefato analisado possui algum tipo de compatibilidade com o catálogo definido, enquanto as não conformidades, indicam alguma incompatibilidade.

Para atingir ao objetivo principal, cada agente trata de forma distinta suas tarefas, recursos, desejos e objetivos. Suas discriminações são apresentadas como:

- MONITOR: o agente monitora constantemente uma pasta onde artefatos devem estar disponíveis para que o SMA consiga efetuar a análise. Após coleta do artefato o Monitor verifica se já há artefatos do agente ANALISADOR referente à memória de avaliações anteriores (base de conhecimento), tais rastros são registrados a partir de uma estratégia de proveniência [38] [39]. A base de conhecimento do ANALISADOR contém o registro de tags das regras de análise baseados na execução do agente Canonizador e a correlação dessas tags com as alternativas do catálogo do RNF de monitorações anteriores.

Dessa forma, caso encontre correlações das informações lidas recentemente com a base de conhecimento, o MONITOR encaminha diretamente ao ANALISADOR as respostas de conformidade sem ter a necessidade de percorrer todo o processo de análise, ou seja, passando por outros agentes como o CANONIZADOR e o CONSOLIDADOR. Dessa forma, tem-se um processo otimizado apoiado pela base de conhecimento formada a partir do histórico de análises. Caso o agente não encontre indicativos de marcações conhecidas e já analisadas anterioemente na base de conhecimento alimentada pelo ANALISADOR ele procede com o processo completo e envia uma mensagem ao agente CANONIZADOR informando sobre a captura e disponibilidade do arquivo. Está definido para o processo otimizado que cem por cento das marcações encontradas no artefato devem ser compatíveis e já analisadas anterioremente.

- CANONIZADOR: tem por objetivo receber os artefatos do MONITOR para proceder com a leitura e comparação estruturas de tags conhecidas, ou seja, aquelas disponíveis em catálogo. Dessa forma, o CANONIZADOR separa artefatos que possuem marcações compatíveis com o catálogo, sem fazer a análise de satisfação do *softgoal* e também artefatos que não possuem marcações ou marcações desconhecidas. Artefatos que possuam apenas marcações compatíveis com variáveis essenciais ou

sinônimos são encaminhados diretamente para o ANALISADOR, apóes pesquisa em base de estruturas conhecidas, enquanto que artefatos com marcações compatíveis com padrões são encaminhadas para o CONSOLIDADOR para que efetue a análise de padrões, ou seja, se o conteúdo disposto sugere compatibilidade com o catálogo.

- CONSOLIDADOR: verifica a partir do catálogo se o conteúdo disponível no artefato, seja delimitado por marcações ou conteúdos avulsos, condiz com estruturas conhecidas e disponibiliza ao ANALISADOR para que possa ser analisado e os resultados obtidos.

- ANALISADOR: efetua a comparação de marcações dos artefatos disponibilizados com o catálogo de RNF. Ao comparar as marcações conhecidas com variáveis essenciais, sinônimos e padrões, o agente inicia o processo de propagação de acordo com *patterns* seleção indicados no catálogo. Com a propagação, o agente sinaliza se o *softgoal* foi satisfeito a contendo e os resultados de conformidade e não conformidades (não atende) são apresentados através de uma trilha que tem origem na alternativa, passa pelas questões e culmina no(s) *softgoal*(s), de acordo com o SIG definido.

A cada análise efetuada o ANALISADOR gera: - base de estruturas conhecidas: armazena em uma estrutura específica as tags que analisou e que estiveram em conformidade com o catálogo. Dessa forma, compõe uma "memória de tags" (assinaturas) já analisadas; - base de conhecimento de proveniência: registra em estrutura própria o rastro de analises dos diversos artefatos. Nesse rastro estão as tags de variáveis essenciais, sinônimos e padrões que satisfizeram alternativas nas análises efetuadas; - resultado a partir da análise da propagação dos elos de relação entre operacionalizações e *softgoals* e entre *softgoals*: registra em estrutura própria o resultado das análises por artefato, conforme poderão ser vistos nos estudos de caso.

A interação dos agentes é determinada pelo modelo intencional conforme Figura 3:

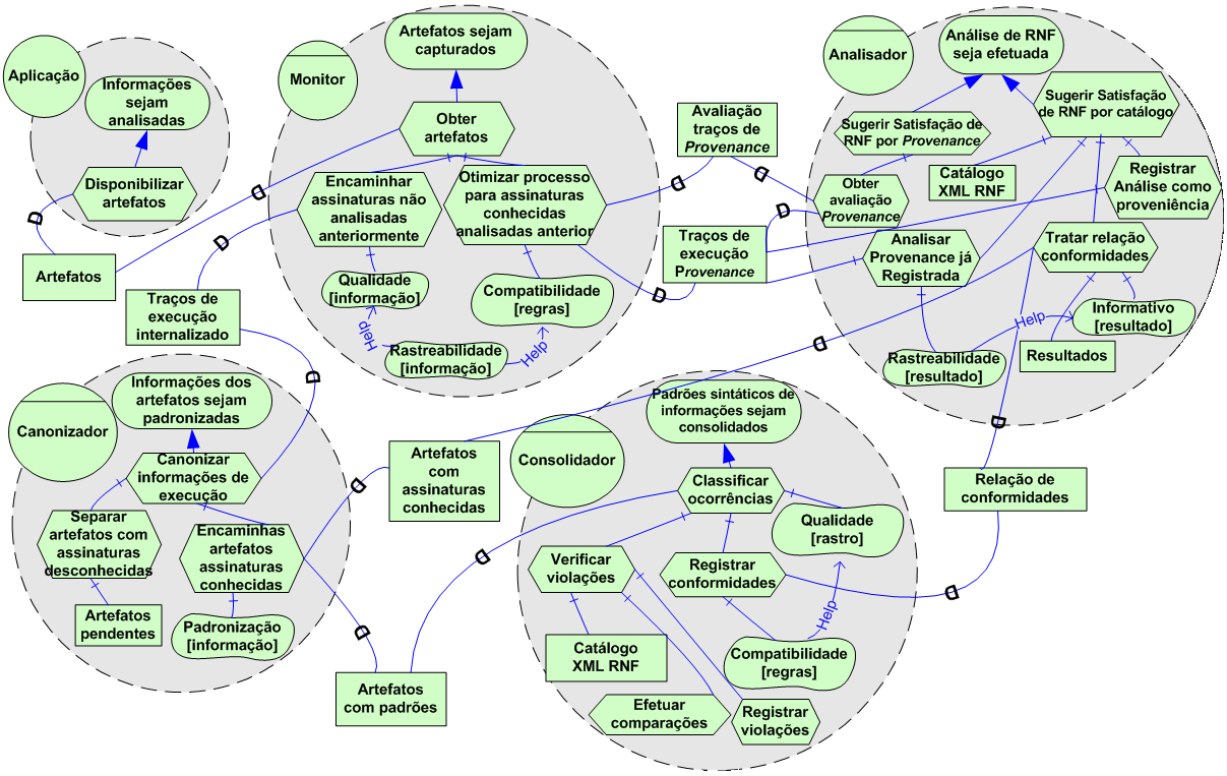


Figura 3. Modelo intencional da interação dos agentes – *framework* i*.

4.5 Estrutura de Assinaturas Conhecidas

As assinaturas conhecidas correspondem àquelas tags que foram analisadas pelo analisador e cujo resultado da análise foi positiva, ou seja, que houve a compatibilidade da tag do artefato, com a tag do catálogo, portanto dada como em conformidade. A estrutura de tags (assinaturas) conhecidas é criada a partir da primeira analise e à medida que outros artefatos são analisados, a estrutura é reestruturada e acrescida de novas tags. Dessa forma, ela só é criada a partir da primeira vez que o SMA é executado.

A estrutura segue o padrão XML e nela são armazenados um identificador, a tag e assinaturas equivalentes, ou sinônimos, registradas no próprio catálogo. Sua composição segue como pode ser apresentado a seguir:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<kbase tipo="RNF" objetivo="estruturas conhecidas">
<KnownStructureBase>
 <KnownStructure>
  <index>0</index>
  <name>agente</name>
  <equivalency>ator<equivalency/>
 </KnownStructure>

 <KnownStructure>
   <index>1</index>
```

```
    <name>data da coleta</name>
    <equivalency>data da gravação<equivalency/>
<equivalency>data do registro<equivalency/>
 </KnownStructure>

 <KnownStructure>
    <index>2</index>
    <name>error</name>
      <equivalency>erro<equivalency/>
     <equivalency>fault<equivalency/>
     <equivalency>falha<equivalency/>
 </KnownStructure>
</KnownStructureBase>
</kbase>
```

Tal estrutura auxilia no processo de análise, uma vez que o CANONIZADOR só permite que sejam processados artefatos que possuem tags conhecidas. Uma vez que isso é identificado, o artefato pode ser analisado pelo ANALISADOR para as devidas comparações com o catálogo para que os resultados sejam obtidos e apresentados.

4.6 Estrutura de Dados de Proveniência

Proveniência (provenance) tem sido utilizado por pesquisadores para manter rastros de dados para a obtenção da forma com que esses dados foram produzidos e para a verificação da qualidade do processo que o produziu [38] [39]. O dicionário Inglês Oxford (2013) [40] define provenance como: '(i) *The fact of coming from some particular source or quarter; origin, derivation*; (ii) *The history of the ownership of a work of art or an antique, used as a guide to authenticity or quality; a documented record of this*'. Pinheiro [41] and Vasquez [42] sugerem que proveniência tem um foco baseado na história, no processo e na transformação da origem dos dados, mas também deveria prover métricas qualitativas e quantitativas.

O objetivo do uso de proveniência nesse trabalho foi o de dar uma solução heurística, que consiste em efetuar uma análise customizada a partir de uma base de conhecimentos de outras análises efetuadas pelo agente ANALISADOR para dar maior agilidade no processo de análise e geração dos resultados. Além disso, a estratégia de proveniência permite o acumulo de conhecimento a partir de diversas análises de software que formam uma base de conhecimento que registraram conformidades com catálogo. A partir dessa abordagem, a proposta de análise pretende estar aderente a um conceito mais amplo de proveniência, que está relacionado não somente na origem da informação, mas também de como ela é derivada, o seu contexto e como é utilizada.

Para prover artefatos de proveniência o agente ANALISADOR mantém um rastro de sua própria execução no momento da comparação das tags dos software analisados com as alternativas, questões e grupos do catálogo de RNF e esse rastro passa a servir como heurística para novas análises.

O recorte superior da Figura 4 aprepresenta uma otimização do processo de analise com o uso da base de conhecimento formada a partir dos traços de proveniência. A interação entre os agentes MONITOR e ANALISADOR requer apenas os artefatos

do software e os de provenance para que seja avaliada e sugerida a conformidade com o RNF, uma vez que o agente MONITOR utiliza os artefatos de provenance como heurística de monitoração e transfere o processamento para o ANALISADOR ao invés de caminhar com o processo para uma varredura mais completa a partir dos agentes CANONIZADOR E CONSOLIDADOR.

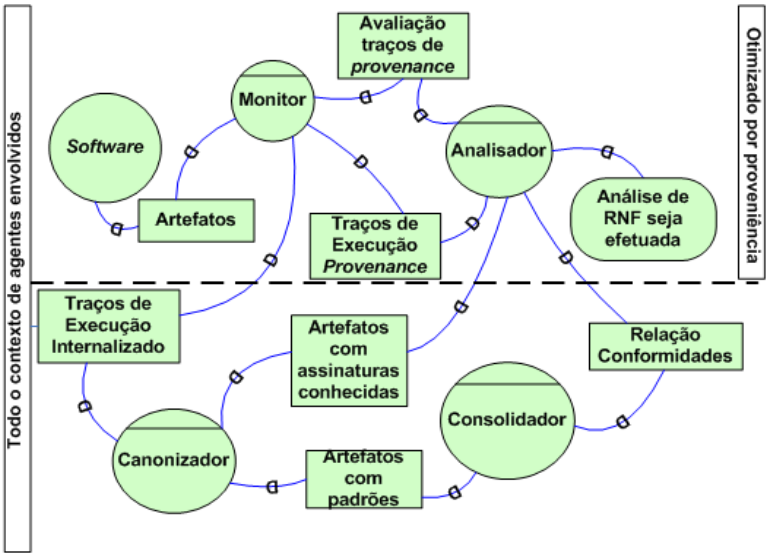


Figura 4. Visão da interação dos agentes com recorte da análise com suporte de proveniência.

O fluxo baseado em registros de provenance evita o uso de todos os agentes no processo de análise. Essa estratégia é baseada na disponibilidade de uma base de conhecimento histórica para dar suporte às decisões do agente MONITOR. Essa base de conhecimento, armazenada a partir de uma abordagem de proveniência, permite ao MONITOR ter acesso às informações de analises anteriores e decidir pela sugestão direta ao agente ANALISADOR que o artefato analisado no momento já possui uma correlação de semelhança com artefatos já analisados.

Os registros de provenance são armazenados em formato XML a partir de cada avaliação do agente ANALISADOR. Os conteúdos são registrados entre as tags, ou seja, à medida que o ANALISADOR encontra uma evidência baseada no catálogo, ele grava o conteúdo encontrado entre as tags indicadoras da conformidade com o catálogo. A estrutura XML criada para armazenar os dados de proveniência segue conforme o código:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<baseproveniencia tipo="RNF" objetivo="proveniência">
    <patternIdentificacao>Transparencia</patternIdentificacao>
      <patternDecomposicaoNivel_1>
        <idDecomposicaoNivel_1>Auditabilidade</idDecomposicaoNivel_1>
            <patternDecomposicaoNivel_2>
                <idDecomposicaoNivel_2>Explicacao</idDecomposicaoNivel_2>
                  <idGrupo>1</idGrupo>
                  <tituloGrupo>Descrever módulos, funções, termos e variáveis utilizadas</tituloGrupo>
                  <patternQuestoes>
                        <idQuestao>1.1</idQuestao>
                        <tituloQuestao>Os módulos, funções, termos e variáveis têm descrições
                                    suficientes?</tituloQuestao>
                        <patternAlternativas>
                              <idAlternativa>1.1.2</idAlternativa>
                              <tituloAlternativa> Utilizar descrições textuais explicativas, baseadas em critério
                                          técnico, ao longo do código fonte, para demonstrar sua relação
                                          com a especificação do software.</tituloAlternativa>
                          <variáveis>
                              <tagsVariável>passodeExecução</tagsVariável>
                          </variáveis>
                        </patternAlternativas>

                  <idGrupo>2</idGrupo>
                  <tituloGrupo>Fornecer Ajuda</tituloGrupo>
                  <patternQuestoes>
                        <idQuestao>2.1</idQuestao>
                        <tituloQuestao>Existem fontes de informação que descrevem o software?
                              </tituloQuestao>
                        <patternAlternativas>
                              <idAlternativa>2.1.1</idAlternativa>
                              <tituloAlternativa> Possuir descrições comentadas com regras condicionais ou de
                                          loop, em linguagem natural estruturada, que corresponda
                                          à implementação de estruturas condicionais ou de
                                          repetição.</tituloAlternativa>
                          <variáveis>
                              <tagsVariável>estruturasAlgoritmo</tagsVariável>
                          </variáveis>
                        </patternAlternativas>
            <patternDecomposicaoNivel_2>
            ...
            <idDecomposicaoNivel_2>Rastreabilidade</idDecomposicaoNivel_2>
                  <id/Grupo>4</idGrupo>
            <tituloGrupo>Fazer Pré-Rastreabilidade</tituloGrupo>
                  <patternQuestoes>
                        <idQuestao>4.1</idQuestao>
                        <tituloQuestao> As fontes de informação utilizadas são rastreadas?</tituloQuestao>
                        <patternAlternativas>
                              <idAlternativa>4.1.2</idAlternativa>
                              <tituloAlternativa> Data em que a mensagem é gravada, registrada ou coletada.
                                          </tituloAlternativa>
                          <variáveis>
                              <tagsVariável>data do registro</tagsVariável>
                          </variáveis>
                        </patternAlternativas>
      </patternDecomposicaoNivel_1>

            ...
</baseproveniencia>
```

Nessa estrutura de base de conhecimento, formada a partir das heurísticas de análise e verificação de artefatos ao longo da execução, o SMA está interessado em armazenar quais tags responderam quais questões a partir do seu vínculo com as

alternativas propostas no catálogo. Portanto, os grupos, questões e alternativas se mantém e tags de variáveis essenciais e sinônimos são adicionados para cada estrutura de grupo, questão e alternativa. A partir do momento em que são analisados diferentes tipos de software comparados ao catálogo, são identificadas quais tags, que historicamente, foram dadas como em conformidade.

5 Arquitetura de implementação dos agentes do SMA

A implementação dos agentes foi baseada em um mapeamento entre o modelo intencional, a arquitetura BDI e a codificação em Jadex [43], conforme proposto em [44]. A Figura 5 apresenta a proposta do mapeamento entre as diferentes camadas.

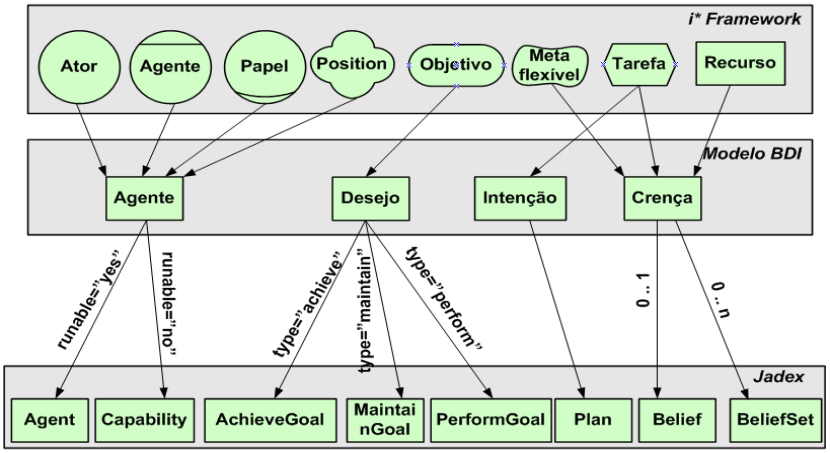


Figura 5. Mapeamento i* x BDI x Jadex [44]

A descrição textual da Figura 5 indica a estratégia de implementação dos agentes. Os agentes não executáveis originados de papéis ou posições deveriam ser implementados como capacidades; os desejos foram traduzidos como objetivos e desenvolvidos ou mantidos de acordo com a tag type do modelo BDI; as intenções foram implementadas com planos, ou seja, classes Java que estenderam os planos da plataforma Jadex [43]; as crenças com cardinalidades de 0 para 1 ou 1 para 1 foram traduzidas para beliefs Jadex, enquanto que crenças com cardinalidades de 0 para n e 1 para n foram traduzidas como conjuntos de beliefset Jadex, conforme proposto em [44]. Em nossa proposta de implementação o CTS foi descrito nas crenças do agente ANALISADOR a partir dos beliefs, que passaram a indicar as questões dos *patterns* de transparência e os beliefset que tiveram o objetivo para indicar as alternativas. Portanto, o SMA desenvolvido não só utiliza das abordagens BDI, como também aplica um mapeamento desde a camada de modelo, abstração em i*, como segue em sua arquitetura de forma organizada ao levar em consideração tal mapeamento.

A especificação BDI dos agentes foi feita em uma arquitetura baseada em XML. O trecho de código da página subsequente apresenta um arcabouço da estrutura do agente ANALISADOR [36] baseado na proposta de [44] do mapeamento i* para BDI e Jadex. Alguns trechos de código foram suprimidos devido à extensão do código

fonte, mas a estrutura procura refletir a arquitetura utilizada na implementação. A implementação visa a refletir as características do método sistêmico apresentadado no modelo intencional da arquitetura dos agentes do SMA. Apesar de funcional, ela ainda é um esforço inicial de codificação para operacionalizar as questões de análise, tal código fonte poderá ser revisto em outro momento para atender a melhores condições de arquitetura e qualidade.

A implementação doa gentes é feita a partir de duas abordagens: a primeira a partir da implementação de estruturas XML para caracterizar os agentes, seus papeis, posições, relações, crenças, desejos e planos; a segunda, a partir de classes java que implementam o comportamento dos agentes. Trechos de código das características da arquitetura dos agentes ANALISADOR e CONSOLIDADOR podem ser vistos como:

- Trecho de código da arquitetura do agente ANALISADOR.

```
<!--
<H3>The Analyzer agent.</H3>
This agent analyze the log for evaluate the RNF
-->
<agent xmlns="http://jadex.sourceforge.net/jadex"
	xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
	xsi:schemaLocation="http://jadex.sourceforge.net/jadex
		http://jadex.sourceforge.net/jadex-0.96.xsd"
	name="AnalyzerAgent"
	package="jadex.examples.TMMAS.agent.analyzerAgent">
	<imports>
		<import>jadex.adapter.fipa.SFipa</import>
		<import>java.io.File</import>
		<import>handleClasses.CanonicalDefinition</import>
		<import>org.w3c.dom.Document</import>
	</imports>
	<beliefs>
		<belief name="stdPath" class="String">
			<fact>"C:/Policy.xml"</fact>
		</belief>
		<belief name="policy" class="Document">
			<fact>null</fact>
		</belief>
	</beliefs>
	<plans>
		<!-- Plan for analyze the log. -->
		<plan name="receiveLog">
			<body class="ReceiveLogPlan"/>
		</plan>

		<!-- Plan for evaluate the RNF according to the operacionalizations. -->
		<plan name="evaluateRNFLog">
			<body class="EvaluateRNFPlan"/>
			<trigger>
				<internalevent ref="call_EvaluateRNFPlan"/>
			</trigger>
		</plan>
  <!-- Plan for provenance registration. -->
```

```
  <plan name="provenanceRegister">
<body class="provenanceRegisterPlan"/>
<trigger>
      <internalevent ref="call_provenanceRegisterPlan"/>
</trigger>
  </plan>
                        <!-- Plan for canonize the different log patterns. -->
                        <plan name="SendRNFEvaluation">
                                <body class="SendRNFEvaluationPlan"/>
                                <trigger>
                                        <internalevent ref="call_SendRNFEvaluationPlan"/>
                                </trigger>
                        </plan>
                        <!-- Plan to save the Unknown Elements. -->
                        <plan name="receiveExecTraces">
                                <parameter name="execTraces" class="Document">
                                <messageeventmapping ref="receiveExecTracesMSG.content"/>
                                </parameter>
                                <body class="EvaluateRNFPlan"/>
                                <trigger>
                                        <messageevent ref="receiveExecTracesMSG"/>
                                </trigger>
                        </plan>
                </plans>

                - Trecho de código da arquitetura do agente CONSOLIDADOR.
<!--
<H3>The Consolidate agent.</H3>
This agent consolidates all the information handled in the process of RNF analyze
-->
<agent xmlns="http://jadex.sourceforge.net/jadex"
                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                xsi:schemaLocation="http://jadex.sourceforge.net/jadex
                        http://jadex.sourceforge.net/jadex-0.96.xsd"
                name="ConsolidatorAgent"
                package="jadex.examples.TMMAS.agent.consolidatorAgent">
                <imports>
                        <import>jadex.adapter.fipa.SFipa</import>
                        <import>handleClasses.CanonicalDefinition</import>
                        <import>handleClasses.OpsOfRNF</import>
                </imports>
                <plans>
                        <!-- Plan for receive the log and register it in the belief base. -->
                        <plan name="ReceiveRaEvCaMo">
                                <body class="ReceiveRaEvCaMoPlan"/>
                        </plan>
                        <plan name="ConsolidateInformation">
                                <body class="ConsolidateInformationPlan"/>
                                <trigger>
                                        <internalevent ref="call_ConsolidateInformationPlan"/>
                                </trigger>
                        </plan>
                </plans>
```

```
        <events>  <messageevent          name="receiveCanonicModelMSG"          type="fipa"
direction="receive">
                        <parameter name="performative" class="String" direction="fixed">
                                <value>SFipa.INFORM</value>  </parameter>
                        <parameter name="content-class" class="Class" direction="fixed">
                                <value>CanonicalDefinition.class</value>
                        </parameter>
                </messageevent>
                <messageevent name="receiveRastEvaluationMSG" type="fipa"
                        direction="receive">
                        <parameter name="performative" class="String" direction="fixed">
                                <value>SFipa.INFORM</value>
                        </parameter>
                        <parameter name="content-class" class="Class" direction="fixed">
                                <value>OpsOfRNF.class</value>
                        </parameter>
                </messageevent>
                <!-- This internal event means the call for the Send RNF Evaluation Plan.-->
                <internalevent name="call_ConsolidateInformationPlan"/>
        </events>
```

A implementação do comportamento dos agentes segue a codificação tradicional em java com definição de classes, métodos... O trecho de código a seguir, representa uma parte de classes criadas para o agente ANALISADOR. Essas classes são derivadas dos planos definidos em XML da arquitetura do agente.

```
package jadex.examples.TMMAS.agent.analyzerAgent;
import jadex.examples.TMMAS.util.handleClasses.CanonicalDefinition;
import jadex.model.IMBelief;
import jadex.model.IMBeliefbase;
import jadex.runtime.IInternalEvent;
import jadex.runtime.IMessageEvent;
import jadex.runtime.Plan;
public class ReceiveLogPlan extends Plan {
        private static final long serialVersionUID = 1L;
        public void body() {
                System.out.println("Analyser: The Analyser agent is ready.");
                System.out.println("Analyser: Waiting for a Canonic log.");
                while (true) {
                  CanonicalDefinition CanonicModel = new CanonicalDefinition();
                  // This will be a new plan. It will receive a message from another
                  // agent and then make something
                  IMessageEvent msg =
              waitForMessageEvent("receiveCanonicLogMSG");
                  System.out.println("Analyser: Log recebido.");
                  CanonicModel = (CanonicalDefinition) msg.getContent();// Get the
              content of the message
                  // Create a new belief in run time to store the canonic log
                  IMBeliefbase model = (IMBeliefbase) getBeliefbase()
                                  .getModelElement();
                  IMBelief belief = model.createBelief("CanonicLog",
                  CanonicalDefinition.class, 0, "false");// Create the belief
                  System.out.println("Analyser: Belief Created.");
```

```
getBeliefbase().registerBelief(belief);// Register belief at runtime
System.out.println("Analyser: Belief Registreated.");
getBeliefbase().getBelief("CanonicLog").setFact(CanonicModel);// Set
        the log in the new belief
// Internal event that represents the reception of the canonic log
IInternalEvent event=
    createInternalEvent("call_EvaluateRNFPlan");
dispatchInternalEvent(event); }  }  }
```

6 Provas de conceito

As provas de conceito do método foi elaborada a partir de sua aplicação em pelo menos quatro realidades de software. O método foi aplicado em software em execução, em código fonte, traços de execução e também em artefatos de arquitetura. Os software escolhidos foram aqueles que os pesquisadores tinham maior confiança em sua execução e na corretude de seus modelos ou artefatos gerados. Os software escolhidos foram:

- Código fonte do Software C&L: O C&L (Cenários e Léxico) [37] é um sistema de software que utiliza a representação de cenários para facilitar a compreensão de domínios da aplicação. O sistema tem como objetivo criar um ambiente colaborativo para criação, edição, manutenção e evolução léxicos e cenários, a partir da disponibilização de informações em linguagem natural semi-estruturada e organizadas de forma simples, de rápida acesso e inter-relacionadas a partir de hipertextos.

- Traços de execução do servidor Apache: Servidor Web livre compatível com protocolo HTTP, compatível com os principais sistemas operacionais como Windows, Linux, Unix e FreeBSD, criado na *National Center for Supercomputing Applications* (NCSA). Em pesquisas de uso, sugere-se que o servidor tenha uma representação de cerca de 47% das plataformas instaladas em 2007 com crescimento para 55%, aproximadamente, em 2010.

- Traços de execução do LS: O LS [24] é um SMA independente para contagem de citações científicas, que combina os serviços do sítio do *Lattes* [25] e do *Google Scholar* [26]. Sua execução é baseada na pesquisa inicial pelo nome do autor no sítio do Lattes e posteriormente uma pesquisa pela relação de artigos no *Google Scholar* com o objetivo de contabilizar citações sobre cada publicação do autor.

- Artefato de arquitetura do LS: A especificação da arquitetura do LS é feito com base no iStarJade[6], que muito se assemelha ao trabalho de Baia et al. (2012) no que diz respeito ao mapeamento entre diagramas i* e a linguagem iStarML [46]. Os agentes tem sua arquitetura implementada a partir de modelos i* com apoio de um padrão de representação textual iStarML [46]. Tal representação é compatível com XML e a partir desse são criadas as estruturas da arquitetura dos agentes.

[6] www.les.inf.puc-rio.br/wiki/images/e/eb/Eduardo_2010_2.ppt

Os estudos foram feitos a partir de dois níveis de *softgoals* do CTS *patternDecomposicaoNivel_1*, composto por Auditabilidade, Acessibilidade, Entendimento e Informativo; e *patternDecomposicaoNivel_2*, por Explicação, Rastreabilidade, Disponibilidade, Detalhamento e Consistência, esses utilizados no material de estudo a partir da exploração de seus *patterns* de nível inferior (grupos, questões e alternativas).

A sistemática da aplicação do método sugere que: se o nível de Decomposição 2 (*patternDecomposicaoNivel_2*) tem seus elementos satisfeitos, os nível imediatamente acima também tem seus elementos satisifeitos. Dessa forma, a propagação segue a mesma cadeia de valores até o nível mais alto da árvore. Portanto, consequentemente Transparência é satisfeita a contento (*satisfice*) [31].

Por motivo da extensão dos estudos será apresentado aqui apenas um detalhamento da prova de conceito realizado sobre os traços de execução do software LS. Maiores detalhes sobre outros estudos podem ser visto em [36].

6.1 Aplicação do Método Sobre Traços de Execução do LS

A aplicação do método no software LS foi feita a partir da análise do RNF de Auditabilidade, definido e presente no CTS (2013) e formalizado em [3]. O CTS traz a definição de Auditabilidade como "Capacidade de ser identificada pela aferição de práticas que implementem características de validade, controlabilidade, verificabilidade, rastreabilidade e explicação" [47].

A aplicação do método pretende avaliar se há no software, objeto da análise, características que satisfazem o requisito de Auditabilidade e uma das possibilidades é no atendimento de um de seus *softgoals* de hierarquia inferior, ou seja, em camada inferior no modelo que é criado. Essa possibilidade pode ser feita para o *softgoal* de Rastreabilidade.

Os dois primeiros estudos basearam-se em artefatos de software em formato ASCii e para o LS serão feitos estudos sobre o seu traço de execução a partir do padrão XML.

As atividades do método foram aplicadas conforme abaixo:

A1 - Criar SIG: Uma série de qualidades é entrada para a A1, nesse caso, Auditabilidade, Validade, Controlabilidade, Verificabilidade, Rastreabilidade e Explicação. Ao final das atividades essas características passam a ser *softgoals* que necessitam ser satisfeitos para atender à Auditabilidade. As relações de Help indicam como são os elos de contribuição entre os *softgoals* folha em relação à Auditabilidade. O resultado final da A1, além da relação dos *softgoals*, é o SIG apresentado na Figura 6.

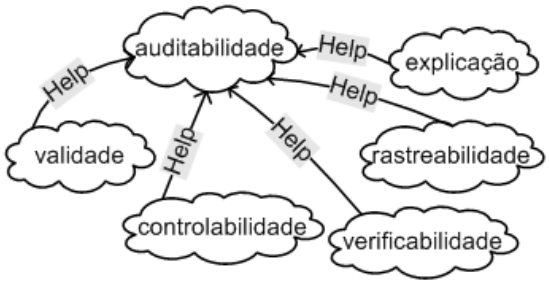


Figura 6. SIG Auditabilidade [3].

O detalhamento de Auditabilidade sugere qualidades que se relacionam com o *softgoal* principal através de um elo de ligação do tipo Help, o que indica que há uma contribuição positiva das qualidades. Para esse estudo é feita uma análise do *softgoal* Rastreabilidade e consequentemente a potencialização de Auditabilidade.

A2 – Definir *patterns:* Para o *Softgoal* Rastreabilidade: A definição de Disponibilidade no CTS corresponde à "Capacidade de seguir a construção ou evolução de uma informação ou de um processo, suas mudanças e justificativas.". A aplicação da A2 para o *softgoal* resulta em um detalhamento de grupos, questões e alternativas de respostas para as questões.

O *softgoal* é decomposto em questões que se respondidas satisfazem à meta definida. Relacionadas às questões as mesmas são agrupadas em grupos que melhor a representam. No caso, Rastreabilidade possui três grupos: 1) Fazer pré-rastreabilidade; 2) Fazer rastreabilidade em tempo de desenho; e 3) Fazer rastreabilidade em tempo de. Os grupos e as questões podem ser vistos conforme apresenta a Figura 7 no quadro Resultado. Tais detalhamentos são propostos no CTS [47] e são definidos a partir da aplicação da A2.1 Descrever questões (descrever questões para *softgoals* folha, no caso Rastreabilidade).

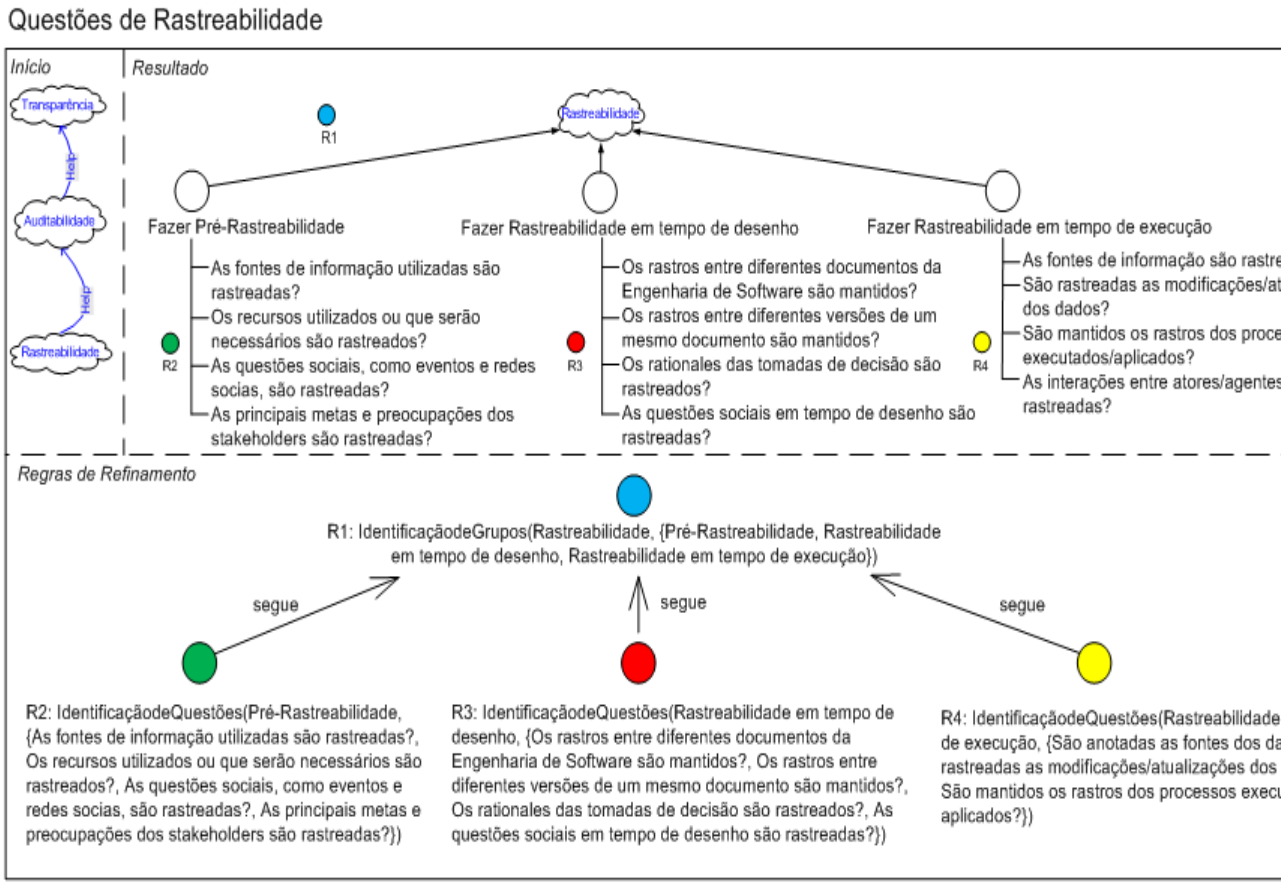


Figura 7. *Patterns* do *softgoal* Rastreabilidade [47].

A decomposição das questões a partir A partir da definição das questões sugere-se composição de respostas alternativas para as questões. Conforme descrito na A2.2 Operacionalizar questões da seção 3.4 Detalhamento de Atividades (Definir *patterns* A2) faz-se necessário o detalhamento das questões em alternativas ou operacionalizações que possam respondê-las. As operacionalizações relatadas a seguir definidas a partir da aplicação da A2.2 foram elaboradas considerando alternativas para disponibilidade segundo o ambiente de execução de aplicação Web. As questão e operacionalização escolhidas aleatoriamente podem ser vistas na Tabela 3 que seguirá a numeração das tabelas descritas na seção anterior. As operacionalizações foram retratadas de forma *ad-hoc* para efeito de estudo da aplicação do método, não necessariamente são estruturas que devem ser consideradas para reuso.

Tabela 3. Relação de grupos, questões e operacionalizações para o *softgoal* Rastreabilidade.

Grupo:	4 Fazer Rastreabilidade em Tempo de Execução.
Questão 1:	4.1 As fontes de informação utilizadas são rastreadas?
Operacionalizações:	4.1.1 São identificados atores que passam a informação
	4.1.2 É identificado o momento em que a informação foi coleta

Questão 2:	4.2 As interações entre atores/agentes são rastreadas?
Operacionalizações:	4.2.1 Registrar troca de mensagens entre os atores/agentes interessados.

A3 – Configurar XML: Para esse estudo serão apresentadas as etapas de A3 e ao final a estrutura do XML.

→Configuração de grupos, questões e operacionalizações:

Definidas as estruturas anteriormente descritas, faz-se necessária sua transposição para uma arquitetura padronizada em XML definida como catálogo XML RNF que servem como regras para os agentes do SMA. Sua transcrição é feita a partir das atividades propostas do detalhamento da A3 Configurar XML. Suas sub-atividades descritas como A3.1 Configurar *Objetive patterns* (estruturas XML), A3.2 Configurar Questions *patterns* (estruturas XML), A3.3 Configurar *Alternative patterns* (estruturas XML) e A3.4 Configurar Variáveis essenciais e seus sinônimos (estruturas XML) são utilizadas para a composição da estrutura XML do catálogo de RNF que servirá como regra para as ações dos agentes do SMA. O XML será apresentado ao final.

→Configuração de variáveis essenciais, sinônimos e padrões:

Após as execuções das atividades 3.1 a 3.3 é necessário executar A3.4 Configurar Variáveis para que se estabeleçam marcadores vinculados às operacionalizações para que as mesmas possam ser identificadas no artefato.

A aplicação da atividade A3.4 Configurar Variáveis essenciais e seus sinônimos é feita inicialmente para o Grupo 1, Questão 1.1 e Operacinalização 1.1.1, bem como a atividade A3.5 Configurar Padrões deram origem às descritas conforme o resultado apresentado a seguir.

Tabela 4. Inserção de Variáveis Essenciais, Sinônimos e Padrões para o *softgoal* Rastreabilidade.

Grupo:	4 Fazer Rastreabilidade em Tempo de Execução.
Questão 1:	4.1 As fontes de informação utilizadas são rastreadas?
Operacionalização:	4.1.1 São identificados atores que passam a informação.
Variável Essencial:	Agente
Noção:	Agente autônomo que interage por troca de mensagens com outros agentes com objetivo de realizar alguma ação.
Sinônimos:	ator, nome do agente, nome do ator, agent name
Padrão:	não se aplica
Operacionalização:	4.1.2 É identificado o momento em que a informação foi coleta
Variável Essencial:	data de gravação
Noção:	Data em que a mensagem é gravada, registrada ou coletada.
Sinônimos:	data da coleta, data de registro
Padrão:	não se aplica

Questão 2:	4.2 As interações entre atores/agentes são rastreadas?
Operacionalizações:	4.2.1 Registrar troca de mensagens entre os atores/agentes interessados.
Variável Essencial:	Conteúdo
Noção:	Conteúdo da mensagem trocada entre agentes.
Sinônimos:	Contente
Padrão:	não se aplica

O catálogo XML RNF para as estruturas definidas até o momento para o estudo, após execução de A3.1, A3.2 e A3.3, segue conforme apresentado a seguir:

```
... (trecho suprimido propositadamente)
<patternDecomposicaoNivel_2>
        <idDecomposicaoNivel_2>Rastreabilidade</idDecomposicaoNivel_2>
        <patternSelecao>HELP</patternSelecao>
        <patternGrupo>
                <idGrupo>4</idGrupo>
                <tituloGrupo> Fazer Rastreabilidade em Tempo de Execução</tituloGrupo>
                <patternQuestoes>
                        <idQuestao>4.1</idQuestao>
                        <tituloQuestao> As fontes de informação utilizadas são
rastreadas?</tituloQuestao>
                                <patternAlternativas>
                                        <idAlternativa>4.1.1</idAlternativa>
                                        <tituloAlternativa> São identificados atores que
passam a
                                                        informação
                                         </tituloAlternativa>
                                        <patternSelecao>RESPONDE</patternSelecao>
                                        <variaveisEssenciais>
                                            <idVariaveisEssenciais>agente
                                           </idVariaveisEssenciais>
                                                <nocaoVariaveisEssenciais>      Agente
autônomo que
                                                    interage por troca de mensagens
com outros
                                                    agentes com objetivo de realizar
alguma ação.
                                                </nocaoVariaveisEssenciais>
                                                <sinonimosVariaveisEssenciais>
                                                 <idSinonimosVariaveisEssenciais>ator
                                                 </idSinonimosVariaveisEssenciais>

<idSinonimosVariaveisEssenciais>nome do
                                                agente</idSinonimosVariaveisEssenciai
s>

<idSinonimosVariaveisEssenciais>nome do ator
                                                 </idSinonimosVariaveisEssenciais>
```

<idSinonimosVariaveisEssenciais>**agent name**
</idSinonimosVariaveisEssenciais>
</sinonimosVariaveisEssenciais>
<padrao>
<idPadrao1>**não se aplica**</idPadrao1>
</padrao>
</variaveisEssenciais>
<idAlternativa>**4.1.2**</idAlternativa>
<tituloAlternativa>**É identificado o momento em que a informação foi coleta**
</tituloAlternativa>
<patternSelecao>**RESPONDE**</patternSelecao>
<variaveisEssenciais>
<idVariaveisEssenciais> **data de gravação**
</idVariaveisEssenciais>
<nocaoVariaveisEssenciais> **Data em que a mensagem é gravada, registrada ou coletada.**
</nocaoVariaveisEssenciais>
<sinonimosVariaveisEssenciais>
<idSinonimosVariaveisEssenciais>**data da coleta**
</idSinonimosVariaveisEssenciais>
<idSinonimosVariaveisEssenciais>**data de registro**
</idSinonimosVariaveisEssenciais>
</sinonimosVariaveisEssenciais>
<padrao>
<idPadrao1>**não se aplica**</idPadrao1>
</padrao>
</variaveisEssenciais>

<idQuestao>**4.2**</idQuestao>
<tituloQuestao> **As interações entre atores/agentes são rastreadas?**
</tituloQuestao>
<idAlternativa>**4.2.1**</idAlternativa>
<tituloAlternativa> **Registrar troca de mensagens entre os atores/agentes interessados.**
</tituloAlternativa>
<patternSelecao>**RESPONDE**</patternSelecao>
<variaveisEssenciais>

<idVariaveisEssenciais>**conteúdo**</idVariaveisEssenciais>
<nocaoVariaveisEssenciais>**Conteúdo da mensagem trocada entre agentes.**

```
                                    </nocaoVariaveisEssenciais>
                                    <sinonimosVariaveisEssenciais>
                                    <idSinonimosVariaveisEssenciais>content
                                    </idSinonimosVariaveisEssenciais>
                                    </sinonimosVariaveisEssenciais>
                                    <padrao>
                                        <idPadrao1>não se aplica</
idPadrao1>
                                    </padrao>
                                    </variaveisEssenciais>

... (trecho suprimido propositadamente)
```

A4 – Configurar Software: A partir dos padrões estabelecidos em catálogo faz-se necessária a configuração do software para que possa gerar estruturas de traços de execução compatíveis com marcadores passíveis de interpretação pelo SMA. Os traços de execução gerados pelo LS são em formato ASCii e divididos em seções. Por exemplo, abaixo é apresentado uma primeira seção gerado pelo Agente1 do LS. Basicamente estão nessa seção as condutas ou ações, baseadas no modelo elaborado para o LS em i*, que estão sendo executadas a cada momento pelo agente.

```
20130913 15:43:24: state=2
20130913 15:43:24: service type=istar.impl.Resource:Citacoes name=scholar:istar.impl.Resource:Citacoes
20130913 15:43:24: service(s) registered
20130913 15:43:25: behaviour type=istar.behaviour.MeansEndUniqueBehaviour name=UrlsBusca adding
subBehaviour
20130913 15:43:25: behaviour type=basicelement.MontarUrlBusca$myBehaviour
name=ObterObrasTopicos
20130913 15:43:25: behaviour type=istar.behaviour.SequentialTaskBehaviour name=ObterCitacoes adding
subBehaviour
20130913 15:43:25: behaviour type=istar.behaviour.MeansEndUniqueBehaviour name=UrlsBusca
20130913 15:43:25: behaviour type=istar.behaviour.MeansEndUniqueBehaviour name=CitacoesPorObra
adding subBehaviour
20130913 15:43:25: behaviour type=basicelement.ObterCitacoesPorObra$myBehaviour
name=ObterCitacoesPorObra
20130913 15:43:25: behaviour type=istar.behaviour.SequentialTaskBehaviour name=ObterCitacoes adding
subBehaviour
20130913 15:43:25: behaviour type=istar.behaviour.MeansEndUniqueBehaviour name=CitacoesPorObra
20130913 15:43:25: behaviour type=istar.behaviour.MeansEndUniqueBehaviour
name=SolitacoesCitacoesAtendidas adding subBehaviour
20130913 15:43:25: behaviour type=istar.behaviour.SequentialTaskBehaviour name=ObterCitacoes
20130913 15:43:25: behaviour type=class maingoal.SolitacoesCitacoesAtendidas$WaitingRequestMessage
name=Initial state=READY
```

A seguir, ao final da primeira seção, são registrados traços de execução da ação do agente na pesquisa por obras, publicações científicas. Nesse ponto, o Agente1 inicia uma preparação de uma base de conhecimento de obras de autores para que possam ser pesquisadas em seguida o número de citações.

```
20130913 15:45:24: Agent name=agente1@VIVALDI:1099/JADE: executing.
Ticket=agente1@VIVALDI:1099/JADE_1
```

20130913 15:45:24: Agent name=agente1@VIVALDI:1099/JADE received a REQUEST for obra: Lexicon Based Ontology Construction.
20130913 15:45:24: Agent name=agente1@VIVALDI:1099/JADE received a REQUEST for obra: Ontology as a Requirements Engineering Product.
20130913 15:45:24: Agent name=agente1@VIVALDI:1099/JADE received a REQUEST for obra: Catologuing Non-Functional Requirements as Softgoal Network.
20130913 15:45:24: Agent name=agente1@VIVALDI:1099/JADE received a REQUEST for obra: Experiences Using Scenarios to Enhance Traceability.
20130913 15:45:24: Agent name=agente1@VIVALDI:1099/JADE received a REQUEST for obra: Um Mecanismo de Rastreamento da Evolução de Cenários Baseado em Transformações.
20130913 15:45:24: Agent name=agente1@VIVALDI:1099/JADE received a REQUEST for obra: Enriquecendo o Código com Cenários.
20130913 15:45:24: Agent name=agente1@VIVALDI:1099/JADE received a REQUEST for obra: Geração de ontologias subsidiada pela Engenharia de Requisitos.
20130913 15:45:24: Agent name=agente1@VIVALDI:1099/JADE received a REQUEST for obra: Indicadores para a Gerencia de Requisitos.
20130913 15:45:24: Agent name=agente1@VIVALDI:1099/JADE received a REQUEST for obra: Domain Networks in the Software Development Process.

Nosso foco de estudo para análise de compatibilidade com os padrões estabelecidos em catálogo de RNF está nessa segunda seção, apesar do LS também gerar outras seções com conteúdos diversos da execução de seus agentes. Para isso faz-se necessário a geração dos traços de execução do LS a partir de um formato XML com tags compatíveis com o catálogo para indicar pontos em que os desenvolvedores do LS tenham inserido características do RNF desejado. Um exemplo de traço de execução para o LS pode ser visto a seguir:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<lattesScholar>
        <mensagem>
                <data do registro>20130913 15:45:24</data do registro >
                <agente> agente1</agente>
                <origem>@VIVALDI:1099/JADE</origem>
                <receiver>server</receiver>
                <destino>@VIVALDI:1099/JADE</destino>
                <tipo>REQUEST for obra</tipo>
                <conteúdo>Ontology as a Requirements Engineering Product</conteúdo>
                <status>received</status>
        </mensagem >
        <menssagem>
                <data do registro>20130913 15:45:24</data do registro>
                <agente> agente1</agente>
                <origem>@VIVALDI:1099/JADE</origem>
                <receiver>server</receiver>
                <destino>@VIVALDI:1099/JADE</destino>
                <tipo>REQUEST for obra</tipo>
                <conteúdo> Catologuing Non-Functional Requirements as Softgoal
                        Network.</conteúdo>
                <status>received</status>
        </messagem>
        <menssagem>
                <data de registro>20130913 15:45:24</data de registro>
                <agente> agente1</agente>
```

```
                    <origem>@VIVALDI:1099/JADE</origem>
                    <receiver>server</receiver>
                    <destino>@VIVALDI:1099/JADE</destino>
                    <tipo>REQUEST for obra</tipo>
                    <conteúdo> Experiences Using Scenarios to Enhance Traceability.</conteúdo>
                    <status>received</status>
          </messagem>
</lattesScholar>
```

A5 – Operacionalizar Agentes: A execução dos agentes se dá pelas atividades A5.1, A5.2 e A5.3, basicamente as atividades representam o funcionamento desde a captura do artefato até sua análise a partir da comparação de marcações ou conteúdos com as regras definidas no catálogo.

O resultado final para o estudo é apresentado em arquivo XML como apresentado a seguir:

Alternativa: 4.1.1 São identificados atores que passam a informação

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<report tipo="RNF" objetivo="ANÁLISE">
    <artefato>agente1-log-2013-09-13 15-45.log </artefato>
    <local>svn/trunk/LattesScholar/file/logs</local>
     <patternIdentificacao>Transparencia</patternIdentificacao>
        <patternDecomposicaoNivel_1>
            <idDecomposicaoNivel_1>Auditabilidade</idDecomposicaoNivel_1>
                <patternDecomposicaoNivel_2>
              <idDecomposicaoNivel_2>Rastreabilidade</idDecomposicaoNivel_2>
                <idGrupo>4</idGupo>
                <tituloGrupo> Fazer Rastreabilidade em Tempo de Execução</tituloGrupo>
                <patternQuestoes>
                            <idQuestao>4.1</idQuestao>
                             <tituloQuestao>  As fontes de informação utilizadas são
rastreadas?</tituloQuestao>
                          <patternAlternativas>
                                <idAlternativa>4.1.1</idAlternativa>
                                <tituloAlternativa> Agente autônomo que interage por troca de
                                  mensagens com outros agentes com objetivo de realizar alguma
ação.
                                 </tituloAlternativa>
                                <situação> EM CONFORMIDADE </situação>
                               <variáveis>
                                   <tagsVariável>agente </tagsVariável>
                                </variáveis>
                                <padrões>
                                    <tagPadrao></tagPadrao>
                                </padrões>
                          </patternAlternativas>
                <patternDecomposicaoNivel_2>
          </patternDecomposicaoNivel_1>
</report>
```

Alternativa: 4.1.2 É identificado o momento em que a informação foi coleta

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<report tipo="RNF" objetivo="ANÁLISE">
    <artefato> agente1-log-2013-09-13 15-45.log</artefato>
    <local> svn/trunk/LattesScholar/file/logs </local>
     <patternIdentificacao>Transparencia</patternIdentificacao>
        <patternDecomposicaoNivel_1>
            <idDecomposicaoNivel_1>Auditabilidade</idDecomposicaoNivel_1>
                <patternDecomposicaoNivel_2>
              <idDecomposicaoNivel_2>Rastreabilidade</idDecomposicaoNivel_2>
                <idGrupo>4</idGrupo>
                <tituloGrupo> Fazer Rastreabilidade  em Tempo de Execução</tituloGrupo>
                <patternQuestoes>
                            <idQuestao>4.1</idQuestao>
                             <tituloQuestao>  As fontes de informação utilizadas são
rastreadas?</tituloQuestao>
                          <patternAlternativas>
                                <idAlternativa>4.1.2</idAlternativa>
                                <tituloAlternativa> Data em que a mensagem é gravada, registrada ou
coletada.
                                </tituloAlternativa>
                                <situação> EM CONFORMIDADE </situação>
                                <variáveis>
                                    <tagsVariável>data de registro </tagsVariável>
                                </variáveis>
                                 <padrões>
                                     <tagPadrao> conteúdo</tagPadrao>
                                 </padrões>
                          </patternAlternativas>
                <patternDecomposicaoNivel_2>
            </patternDecomposicaoNivel_1>
</report>
```

Alternativa: 4.2.1 Registrar troca de mensagens entre os atores/agentes interessados

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<report tipo="RNF" objetivo="ANÁLISE">
    <artefato> agente1-log-2013-09-13 15-45.log</artefato>
    <local> svn/trunk/LattesScholar/file/logs</local>
     <patternIdentificacao>Transparencia</patternIdentificacao>
        <patternDecomposicaoNivel_1>
            <idDecomposicaoNivel_1>Auditabilidade</idDecomposicaoNivel_1>
                <patternDecomposicaoNivel_2>
              <idDecomposicaoNivel_2>Rastreabilidade</idDecomposicaoNivel_2>
                <idGrupo>4</idGrupo>
                <tituloGrupo> Fazer Rastreabilidade em Tempo de Execução </tituloGrupo>
                <patternQuestoes>
                            <idQuestao>4.2</idQuestao>
                             <tituloQuestao>  As interações entre atores/agentes são
rastreadas?</tituloQuestao>
                             <patternAlternativas>
                                <idAlternativa>4.2.1</idAlternativa>
```

```
                    <tituloAlternativa>Registrar troca de mensagens entre os
atores/agentes
                                                    interessados.
                    </tituloAlternativa>
                    <situação> EM CONFORMIDADE </situação>
                    <variáveis>
                        <tagsVariável>conteúdo </tagsVariável>
                    </variáveis>
                    <padrões>
                        <tagPadrao></tagPadrao>
                    </padrões>
                </patternAlternativas>
            <patternDecomposicaoNivel_2>
        </patternDecomposicaoNivel_1>
</report>
```


O estudo realizado contempla a análise dos traços de execução do LS e permite uma comparação automática de tags internas com o catálogo de RNF. Nesse estudo há também a necessidade da avaliação por interação humana para indicar se os conteúdos das tags correspondem com o que sugere a tag. Tal análise de conteúdos para estruturas de maior volume de registros não está contemplada no SMA, uma vez que seriam necessárias estruturas de maior complexidade em substituição às estruturas de tags do tipo <padrão>. Apesar disso, o SMA identifica artefatos gerados pelos traços de execução e informa se há nesse artefato características que condizem com o catálogo pré-estabelecido. Caso não sejam encontradas evidências (conformidades) de tags compatíveis com o catálogo, o SMA registra tal ocorrência e informa que o artefato não está em conformidade (<situação>NÃO ATENDE</situação>) às regras estabelecidas para as alternativas registradas.

A análise a partir da interação humana do LS indica que tal software está em conformidade com as regras estabelecidas para o *softgoal* de rastreabilidade no que se refere à partes do grupo Fazer Rastreabilidade em Tempo de Execução.

7 Conclusão

O presente trabalho de pesquisa apresentou uma proposta de método sistêmico para análise de RNFs definidos em catálogo que une abordagem de GORE à analise automática ou semi-automática a partir de operacionalizações efetuadas por SMA com agentes definidos pelo modelo BDI.

Aspectos conceituais foram considerados a partir de teorias de GORE, RNF *framework*, GQM, GQO e RNF *patterns* que dão suporte teórico ao que é construído enquanto método sistêmico proposto no trabalho; enquanto que para a construção da aplicação e sua as arquiteturas físicas suportadas pelo SMA estão fundamentos sobre modelagem intencional, *framework* i*, painel de intencionalidades, SMA e BDI. Dessa forma, entende-se que o trabalho possuim uma prática de construção TOP-DOWN uma vez que inicia a partir de definição de modos conceituais de determinado RNF, bem como a partir de intencionalidade de agentes envolvidos no SMA, até às definições de camadas físicas de mais baixo nível da aplicação computacional. Tal abordagem

também é utilizada pela definição do catálogo utilizado, em [47], para a definição de mais alto nível de *softgoals* que satisfação a um determinado RNF analisado até o mais baixo nível, onde estão operacionalizações funcionais que satisfazem aos níveis conceituais acima. Tratar RNFs a partir de modelos GORE não é uma nova frente de pesquisa, porém a aplicação de GORE aliada a *patterns* operacionalizados por SMA para análise de satisfação é uma novidade na literatura, principalmente quando esses *patterns* compõem um catálogo que alia a modelagem conceitual às alternativas (operacionalizações) de respostas funcionais que possam satisfazer à RNFs.

Essas abordagens unidas em um método sistematizado, ou seja, que deixe claro seus passos metodológicos baseados em critérios técnicos, compõem um desafio, uma vez que é necessária o estudo e criação de uma sistemática que leve em consideração, desde a concepção de modelos conceituais, até arquiteturas físicas para implementação de uma aplicação que possa lidar com a verificação de RNF que tem por característica um elevado grau de subjetividade dependente do ponto de vista de quem o avalia. Isso foi viabilizado a partir da criação de uma estrutura (catálogo) que associasse, a partir da modelagem proposta no NFR *framework* [4] e GQO [13], metas flexíveis a operacionalizações a partir de notações XML que pudessem ser interpretadas e verificadas condições de satisfação por agentes autônomos.

Alguns trabalhos futuros são vislumbrados:

- O foco inicial da análise foi a verificação de conformidades dos registros em relação aos patterns de RNF, apesar disso o SMA permite o armazenamento de não conformidades o que pode ser explorados com a proposta do uso de sistemas de recomendação [48] [49]. A intenção é criar agentes autônomos inteligentes que possam sugerir alternativas a partir do catalogo de RNF para que as não conformidades encontradas sejam mitigadas. Para isso as estratégias de proveniência e bases de conhecimento devem ser reestruturadas para agentes possam analisar ações pré-estabelecidas e ações executadas para que se minimize o desvio entre planejado e realizado.

- Extender o catálogo XML para que incorpore estruturas formais que permitam análise sintática e semântica baseadas em ontologias sem que percam seu grau de valor absoluto que lhes permitem ser analisáveis. Isso poderia minimizar ou extinguir a necessidade de interação humana em analises feitas, por exemplo, a partir das atuais variáveis essenciais; Aplicar o método sistêmico de análise de RNF para criar catálogos que possam ser aderentes a modelos de maturidade ou ciclos de vida de software como CMMI, MPS.BR ou ISO 12207. As estruturas criadas poderiam contemplar catálogos que atendessem a processos de gestão de requisitos ou gestão da configuração, para que o SMA analisasse se artefatos gerados pelos processos estão condizentes com normas estabelecidas e atendem as exigências dos modelos de maturidade;

- Na área de agentes, um campo de pesquisa tem sido a associação de normas em SMA com o objetivo de regular o comportamento de agentes sem que os mesmos percam sua autonomia. As normas podem ser definidas por restrições, responsabilidades ou padrões de comportamento para atingir determinado objetivo, a

fim de que sejam satisfeitos ou mesmo evitados [50] [51]. O catálogo XML de RNF pode compor tais normas para que regras de análise sejam postas a fim de que os agentes operem baseado em contexto pré-estabelecido a partir de uma abordagem NBDI [50]. Apesar de ter uma semelhança do que é proposto no trabalho de tese, o mesmo não formaliza tal pesquisa baseada nos pressupostos teóricos de NBDI e não trata de questões de restrição, como por exemplo poderia ser aplicado numa camada de Padrões Alternativas/Problemas, em que problemas poderiam ser tratados como regras de restrição.

Contribuição dos autores:

- André Castro: O autor participou na coleta de dados, análise bibliográfica, elaboração de modelos, definição de arquiteturas, definição do método e na aplicação prática das provas de conceito.

- Henrique Sousa: O autor participou das reuniões de pesquisa para entendimento da elaboração de algoritimos necessários para verificação automática de modelos.

- Julio Leite: O autor participou diretamente na orientação do trabalho de pesquisa.

Referências

1. Fenton, N. E.; Pfleeger, S.L.. Software Metrics: a Rigorous and Practical Approach, 2nd Ed., PWS Publishing Company. 1997.
2. Leite, J.C.S.P.; Cappelli, C.. Software Transparency. Business & Information Systems Engineering, Springer, 2010, pp 127-139.
3. Cappelli, C.. Uma Abordagem para Transparência em Processos Organizacionais Utilizando Aspectos. Rio de Janeiro. 328 p. Tese de Doutorado – Departamento de Informática, PUC-Rio, 2009.
4. Chung, L.; Nixon, B. A.; YU, E., and Mylopoulos, J. Non-Functional Requirements in Software Engineering. Springer, 2000.
5. Sommerville, I.. Engenharia de Software, 9ª Edição. Pearson Education, 2011.
6. Fidge, C.; Lister, A.. The Challenges of Non-Functional Computing Requirements. Seventh Australian Software Engineering Conference (ASWEC93), 1993.
7. Matoussi, A.; Laleau, R.. A survey of Non-Functional Requirements in Software Development Process, Universita Paris, Faculté des Science et Technologie, Paris 2008
8. Kaiya, H.; Kaijiri, K... Refining Behavioral Specification for Satisfying Non-functional Requirements of Stakeholders. In IEICE transactions on information and systems Vol.E85-D, No.4(20020401), pages 623–636, 1999.
9. Glinz, M. On Non-Functional Requirements. In 15th IEEE International Volume , Issue , 15-19 Oct, pages 21–26, 2007.

10. van Lamsweerde, A.. Goal-Oriented Requirements Engineering: A Guided Tour. 5th IEEE International Symposium on RE'01, pp. 249-262, August 2001.

11. van Lamsweerde , A.; Letier, E.. Handling Obstacles in Goal-Oriented Requirements Engineering. IEEE Trans. Software Eng., vol. 26, pp. 978-1005, 2000.

12. Supakkul, S.; Hill, T; Chung, L.; Than Tun, T.; Leite, J.C.S.P.. An NFR Pattern Approach to Dealing with NFRs. In: 18th IEEE International Requirements Engineering Conference, 2010, Sydney. Procceedings of the 18th IEEE International Requirements Engineering Conference. los alamitos : ieee computer society press, 2010. v. 18. p. 179-188.

13. Serrano, M.; Leite; J.C.S.P.. Capturing transparency-related requirements patterns through argumentation. In: First International Workshop on Requirements Patterns (RePa), pp.32-41, 29 Aug. 2011.

14. Basili, V. R.. Software Modeling and Measurement: The Goal Question Metric Paradigm. Computer Science Technical Report Series, CS-TR-2956 (UMIACS-TR-92-96), University of Maryland, College Park, MD, September 1992.

15. Rumbaugh, J.; Blaha, M.; Premerlani, W.; Eddy, F.; Lorensen, W.. Object-Oriented Modeling and Design. Prentice Hall. 1991.

16. Demarco. T., Structured Analysis and System Specification. Yourdon Press. 1978.

17. Ross, D.. Structured Analysis (AS): A Language for Communicating Ideas (SADT). IEEE Transactions on Software Engineering, vol. 3, no.1, 1977, pp. 16-34.

18. Lapouchnian, A.. Goal-Oriented Requirements Engineering: An Overview of the Current Research. Technical report, Universidade de Toronto, Canadá, 2005.

19. Bratman, M. E. Intention, Plans, and Practical Reason. University of Chicago, ISBN: 1575861925, 208 pages, 1999.

20. Braubach, L.; LAMERSDORF, W.; POKAHR, A.. Jadex: Implementing a BDI Infrastructure for JADE Agents. Distributed Systems and Information Systems, vol. 3, n. 3, pp.76-85, September 2003.

21. Rao, A. S.. AgentSpeak: BDI agents speak out in a logical computable language, in 'MAAMAW '96: Proceedings of the 7th European workshop on Modelling autonomous agents in a multi-agent world: agents breaking away', Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996, pp. 42–55.

22. Leal, A. L. C.; Sousa, H. P; Leite, J. C. S. P; Lucena, J. P.. Aplicação de modelos intencionais em sistemas multiagentes para estabelecer políticas de monitoração de transparência de software. Revista de Informática Teórica e Aplicada, vol. 20, no. 2, 2013, pp. 111-138.

23. Leal, A. L. C.; Sousa, H. P.; LEITE, J. C. S. P.. Desafios de monitoração de requisitos não funcionais: avaliação em Transparência de Software. In: Requirements Engineering@Brazil 2013, 2013, Rio de Janeiro. CEUR Workshop Proceedings, 2013. v. 1005. p. 25-30.

24. Lattesscholar: Requirements Engineering Group at PUC-Rio. <Disponível em: http://www.er.les.inf.puc-rio.br/~wiki/index.php/Lattesscholar> <Acessado em: 12/05/2015>

25. Lattes. Plataforma Lattes: Currículo Lattes. <Disponível em: lattes.cnpq.br > <Acessado em: 12/05/2015>

26. Google Scholar. <Disponível em: www.scholar.google.com > <Acessado em: 12/05/2015>
27. Publish and Perish. <Disponível em: http://www.harzing.com/pop.htm> <Acessado em: 12/05/2015>
28. Microsoft Academic Search. <Disponível em: http://academic.research.microsoft.com> <Acessado em: 12/05/2015>
29. Google Scholar Citation <Disponível em: http://scholar.google.com/intl/en/scholar/citations.html> <Acessado em: 12/05/2015>
30. Yu, E.S.K.. Modelling Strategic Relationships For Process Reengineering. Ph.D. dissertation. Dept. of Computer Science, University of Toronto, 1995.
31. Cunha, H.S.. Uso de estratégias orientadas a metas para modelagem de requisitos de segurança. Rio de Janeiro, 2007. 145p. Dissertação de Mestrado - Departamento de Informática, PUC-Rio.
32. Oliveira, A.P.A.; Leite, J.C.S.P.; Cysneiros, L. M.. Engenharia de Requisitos Intencional: Um Método de Elicitação, Modelagem e Análise de Requisitos. Rio de Janeiro, 2008. 261 p. Tese de Doutorado - Departamento de Informática, PUC-Rio.
33. Simon H. A.. The Sciences of the Artificial. MIT Press, Cambridge, MA, USA, 3rd ed., 1996.
34. Leite, J. C. S. P.; Rossi, G.; Maiorana, V.; Balaguer, F.; Kaplan, G.; Hadad, G.; Oliveros, A.. Enhancing a requirements baseline with scenarios. In: IEEE INTERNATIONAL SYMPOSIUM ON REQUIREMENTS ENGINEERING – RE97, 3rd, 1997, Annapolis, MD. Proceedings. IEEE Computer Society Press, 1997. p. 44-53.
35. Wooldridge, M.. An Introduction to Multi-Agent Systems. John Wiley and Sons. 2002.
36. Leal, A. L. de C.. Análise de Conformidade de Software com Base em Catálogos de Requisitos não Funcionais: Uma Abordagem Baseada em Sistemas Multi-Agentes. Rio de Janeiro. 196 p. Tese de Doutorado – Departamento de Informática, PUC-Rio, 2014.
37. Leite, J.C.S.P.; Hadad, G.; Doorn, J.. Kaplan, G.. A Scenario Construction Process. In.: Requirements Engineering Journal. Springer-Verlag London Limited: 2000, Vol. 5, n.1, Pags. 38-61.
38. Miles, S.; Groth, P.; Branco, M.; Moreau, L.. The requirements of recording and using provenance in e-Science experiments. In.: Technical Report: Electronics and Computer Science, University of Southampton. 2005.
39. Miles, S.; Munroe, S.; Luck, M.; Moreau, L.. Modelling the provenance of data in autonomous systems. In.: Proceedings of Autonomous Agents and Multi-Agent Systems 2007, pp. 243–250, Honolulu, Hawai'i, May.
40. Oxford English Dictionary, "provenance, n". Oxford University Press.
41. Pinheiro da S.; MCGUINNESS, P.; MCCOOL, D.. Knowledge Provenance Infrastructure. IEEE Data Engineering Bulletin 26(4), 2003, pp. 26-32.
42. Vasquez I.; Gomadam K.; Patterson S.. Framework for representing provenance for web services and processes. Technical Report, LSDIS Lab, 2005.
43. Braubach, L.; Lamersdorf, W.; Pokahr, A.. Jadex: Implementing a BDI Infrastructure for JADE Agents. Distributed Systems and Information Systems, vol. 3, n. 3, pp.76-85, September 2003.

44. Serrano, M.; Leite, J. C. S. P.. Development of Agent-Driven Systems: from i* Architectural Models to Intentional Agents Code. In: Fifth International istar Meeting, 2011, Itália. Fifth International istar Meeting.

45. Leite, J. C. S. P. ; Yu, Y ; Liu, L. ; Yu, E. ; Mylopoulos, J . Quality-Based Software Reuse. In: Advanced Information Systems Engineering: 17th International Conference, CAiSE 2005, 2005, Porto, Portugal. Advanced Information Systems Engineering: 17th International Conference, CAiSE 2005, Porto, Portugal, June 13-17, 2005. Proceedings, 2005. v. 17. p. 535-545.

46. Cares, C.; Franch, X.; Perini, A.; Susi, A.. iStarML: The i* Mark-up Language: References Guide. Barcelona, Spain, August 2007.

47. CTS. Catálogo de Transparência de Software. 2013. <Disponível em: http://transparencia.inf.puc-rio.br/wiki/index.php/Cat%C3%A1logo_Transpar%C3%AAncia> <Acessado em: 17/05/2015>.

48. Adomavicius, G.; Tuzhilin, A.. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering, Vol. 17(6), 2005, pp. 734-749.

49. Burke, R.. Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction, Vol. 12(4), 2002, pp. 331-370.

50. Neto, B. F. dos S.; da Silva, V. T.; de Lucena, C. J. P.. NBDI: An architecture for goal-oriented normative agents. In.: ICAART 2011 - Proceedings of the 3rd International Conference on Agents and Artificial Intelligence, Volume 1 – Artificial Intelligence, Rome, Italy, January 28-30, 2011, pp. 116–125.

51. Camino, A. G.; Normative regulation of open multi-agent systems, Ph.D. dissertation, Artificial Intelligence Research Institute (IIIA), Spain, 2009.

Explorando mapas de relacionamento com base em subtópicos para sumarização multidocumento

Exploring the subtopic-based relationship map strategy for multi-document summarization

Rafael Ribaldo [1]
Paula Christina Figueira Cardoso [2]
Thiago Alexandre Salgueiro Pardo [3]

Data de submissão: 04/10/2015, *Data de aceite: 04/04/2016*

Abstract: Neste artigo, foi adaptada e explorada uma estratégia de geração de sumários multidocumento com base na abordagem de mapas de relacionamento, a qual representa textos como grafos (mapas) de segmentos inter-relacionados e aplica técnicas de percurso em grafo para produzir os sumários. Em particular, utilizou-se o Caminho Denso-Segmentado, um método sofisticado que tenta representar em um sumário os principais subtópicos dos textos de origem, mantendo a informatividade do sumário. Além disso, também foram investigadas algumas técnicas de segmentação e agrupamento de subtópicos bem conhecidas, a fim de selecionar corretamente as informações mais relevantes para compor o sumário final. Mostra-se que este método de sumarização baseado em subtópicos supera outros métodos de sumarização multidocumento e atinge resultados do estado da arte, competindo com os métodos de sumarização profundos mais sofisticados da área.

***P*alavras-chave:** sumarização multidocumento, mapas de relacionamento, grafos, subtópicos, segmentação e agrupamento

[1]Núcleo Interinstitucional de Linguística Computacional (NILC), Departamento de Ciências de Computação, Instituto de Ciências Matemáticas e de Computação, USP,São Carlos/SP, Brasil. {ribaldorafael@gmail.com}
[2]Departamento de Ciência da Computação, Universidade Federal de Lavras, Avenida Doutor Sylvio Menicucci, 1001, Campus Universitário. CEP: 37200-000 - Lavras/MG, Brasil. {paula.cardoso@dcc.ufla.br}
[3]Núcleo Interinstitucional de Linguística Computacional (NILC), Departamento de Ciências de Computação, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos/SP, Brasil. {taspardo@icmc.usp.br}

Abstract: In this paper we adapt and explore a strategy for generating multi-document summaries based on the relationship map approach, which represent texts as graphs (maps of interrelated segments and apply traversing techniques for producing the summaries. In particular, we work on the Segmented Bushy Path, a sophisticated method which tries to represent in a summary the main subtopics from the source texts while keeping its informativeness. In addition, we also investigate some well-known subtopic segmentation and clustering techniques in order to correctly select the most relevant information to compose the final summary. We show that this subtopic-based summarization method outperforms other methods for multi-document summarization and achieves state of the art results, competing with the most sophisticated deep summarization methods in the area.

***K*eywords:** multi-document summarization, relationship maps, graphs, subtopics, segmentation and clustering

1 Introduction

In recent decades, many new technologies have emerged, bringing with it an increasing volume of information. Nowadays, many resources such as search engines, blogs and social networks make accessible an enormous amount of information and, therefore, processing all this becomes increasingly difficult. To have an idea, a study conducted by the International Data Corporation (IDC) showed that the volume of digital content would grow to 8 ZB (zettabytes) in the last year, driven by steady growth of internet users, social networks and smart devices, which enable new ways of working and communicating. In this context, multi-document summarization (MDS) appears as a tool that may assist people in acquiring relevant information in a short time.

Multi-document summarization aims at producing automatic summaries from a collection of source texts/documents about the same topic [1]. Some challenges for MDS are to deal with the multi-document phenomena, as redundant, complementary and contradictory information, to normalize varied writing styles (since the texts come from different authors), to temporally order events/facts (because the texts are written at different times), to balance different foci and perspectives, as well as to keep summary coherence and consistency. It also includes the traditional single document summarization challenges, as dealing with dangling anaphors and guaranteeing cohesion.

An equally important challenge for MDS is to tackle the topic/subtopic distribution in summaries. It is known that a text or a set of texts develop a main topic, exposing several subtopics as well. A topic is a particular subject that we write about or discuss, and subtopics are represented in pieces of text that cover different aspects of the main topic [2, 3, 4, 5]. A good summary should represent the main topics/subtopics of the source texts in order to be

considered relevant and informative to the reader.

As an example, Figure 1 shows two texts (translated from the original language - Portuguese) and possible subtopic delimitations. Sentences are identified by numbers between square brackets. The identification of each subtopic is shown in angle brackets after the corresponding passage. As it is possible to see, the main topic is the health of Maradona, the famous Argentine soccer player (already retired). Notice that it is possible to find subtopics about the disease itself and the soccer match between Boca Juniors and River Plate teams.

Document 1	$[S1]$Maradona's personal doctor, Alfredo Cahe, revealed on Monday that a relapse of acute hepatitis that the ex-player suffers was the reason for his new hospitalization. $[S2]$Maradona had been discharged 11 days ago, but he was again hospitalized on Friday, and the medical reports did not specify what was wrong with the ex-player — Cahe ruled out ulcer or pancreatitis. *<subtopic: Maradona's relapse>* $[S3]$"Maradona had a relapse of acute hepatitis. Now he is stable. Despite the fact that he got better on Sunday, he should remain hospitalized", said Cahe to the newspaper La Nación. *<subtopic: Maradona's hepatitis>* $[S4]$Maradona, 46, developed toxic hepatitis due to excessive alcohol consumption, which had kept him hospitalized for 13 days before the most recent hospitalization. *<subtopic: history of Maradona's disease>* $[S5]$Cahe said that Maradona had not started to drink alcoholic beverages again, and that the causes of the relapse are being investigated. *<subtopic: Maradona's hepatitis>*
Document 2	$[S1]$Maradona had again health problems over the weekend. $[S2]$Hospitalized in Buenos Aires, he had a relapse and felt pain againg due to acute hepatitis, according to his personal doctor, Alfredo Cahe. *<subtopic: Maradona's relapse>* $[S3]$"Now his state of health is stable. Despite this improvement, he is still hospitalized", said the doctor, who has discarded the possibility that the ex-player has pancreatitis (inflammation of the pancreas, an organ located behind the stomach and that influences the digestion). $[S4]$Cahe emphasized that Maradona still has problems. $[S5]$"His liver values are not balanced and he is not well. But it is nothing serious", he said in an interview for the La Nación newspaper. *<subtopic: current state of health>* $[S6]$On Sunday, Maradona watched the 1-1 draw in the classic Boca Juniors and River Plate on television. $[S7]$Boca Juniors' fans, who turned out in large numbers to the stadium La Bombonera, led many banners and flags with messages of support for the Argentine idol. $[S8]$His daughter, Dalma, was in the stadium to watch the game. *<subtopic: support messages>*

Figure 1. Example of documents with diverse information

Considering the multi-document scenario, it is easy to find repeated subtopics in the

texts, for example, <*subtopic: Maradona's relapse*>, and also unique subtopics that are not repeated and contain different details about the main topic, for example, <*subtopic: messages of support*>. In this case, before selecting content for a summary, it is necessary to find similar subtopics and clustering them according to a degree of similarity. Ideally, a multi-document summary should contain only once the key shared information among all the documents, plus other information unique to some individual documents that show to be relevant [6].

As an illustration of the necessity of subtopic treatment, we show in Figures 2 and 3 two summaries (also translated from Portuguese) automatically produced by GistSumm [7] and CSTSumm [8] systems, which use superficial and deep summarization methods, respectively. Superficial methods are those that make little or no use of linguistic knowledge, and are more scalable and robust in general. Deep methods, by contrast, make heavy use of linguistic knowledge, such as discourse models and semantic resources, being able to produce better results, but are more expensive and more narrowly applicable, typically. GistSumm system uses simple word frequency measures to identify the gist (the main idea) of the texts. CSTSumm system, in turn, uses discourse features for judging sentence relevance. We may notice that both summaries do not include all the subtopics present in the collection of texts, since the summarization strategies used by GistSumm and CSTSumm make uniform use of the sentences in different documents [9], i.e., the sentences are used without consideration of the subtopic distribution. We and also other authors [9, 10, 11] argue that sentences in the same collection may not be uniformly treated, because some sentences are more important than others, due to their different roles in the documents and the subtopics they belong to.

"Maradona had a relapse of acute hepatitis. Now he is stable. Despite the fact that he got better on Sunday, he should remain hospitalized", said Cahe to the newspaper La Nación.
"Now his state of health is stable. Despite this improvement, he is still hospitalized", said the doctor, who has discarded the possibility that the ex-player has pancreatitis (inflammation of the pancreas, an organ located behind the stomach and that influences the digestion).

Figure 2. Summary produced by GistSumm system

Hospitalized in Buenos Aires, he had a relapse and felt pain again due to acute hepatitis, according to his personal doctor, Alfredo Cahe.
"Maradona had a relapse of acute hepatitis. Now he is stable. Despite the fact that he got better on Sunday, he should remain hospitalized", said Cahe to the newspaper La Nación.
Cahe said that Maradona had not started to drink alcoholic beverages again, and that the causes of the relapse are being investigated.

Figure 3. Summary produced by the CSTSumm system

In order to overcome the limitations of the current summarization strategies, in this paper we adapt and explore a classical summarization technique proposed by Salton et al. [3]. The authors have proposed single document summarization methods that are referred by "relationship maps", since a text is represented as a graph (a "map") of interrelated text segments and different traversing techniques are used to select the segments to compose a summary. We have already adapted two of the methods (called "bushy" and "depth-first" paths) for multi-document summarization (as reported in [12]). Here, we adapt and explore the most sophisticated method, the "Segmented Bushy Path", which addresses the subtopic issues for producing better summaries. As the Segmented Bushy Path was developed for single document summarization, it is necessary more than segmenting each text in subtopics for adapting it to MDS: it is also necessary to deal with subtopic correlations. For this reason, this paper also deals with strategies for subtopic segmentation and clustering.

We evaluate the adapted method in a benchmark collection for Portuguese language and show that it outperforms the other Salton et al. methods [3] and that it produces state of the art results, competing with the best deep method that we are aware of. We also comment on the delivering of our methods to the general public by incorporating them in an extension to a web browser, making the summarization system widely available.

The remainder of the paper is structured as follows. In the next section, we briefly review the main related work. In Section 3, we present the summarization algorithm and its steps. Experiments and evaluation results are reported in Section 4. In Section 5, we briefly describe our initiative to deliver to the final user our summarization methods. Finally, some final remarks are presented in Section 6.

2 Related Work

In what follows, we introduce the main related work and concepts that are the basis of this paper. We briefly review some related summarization methods already tested for Portuguese and mainly the ones of Salton et al. [3], which support our work, followed by a discussion of subtopic segmentation and clustering techniques. We conclude this section with a description of a discourse model that we test in this paper for subtopic clustering.

2.1 Text Summarization

Graphs have shown to be applicable to many Natural Language Processing applications [13] and there are several graph-based approaches for both single and multi-document summarization (see, for example, [3, 8, 9, 14, 15, 16, 17, 18, 19]).

Salton et al. [3] probably introduced the first widespread graph-based approach to single document summarization. In the proposed relationship map methods, the authors model

a text as a graph/map in the following way: each paragraph is represented as a node, and weighted links are established only among paragraphs that have some lexical similarity. This may be pinpointed through lexical similarity metrics. The choice for representing paragraphs (and not words, clauses, or sentences) as nodes is due to the assumption that paragraphs provide more information surrounding their main topics and, thus, may be used for producing more coherent and cohesive summaries. For summarization purposes, only the highly weighted links are considered: given a graph with N nodes, only the $1.5 * N$ best links provide the means to select paragraphs to include in a summary. Once the graph is built, three different ways of traversing the graph are proposed, namely, Bushy Path, Depth-first Path and Segmented Bushy Path. In the Bushy Path, the density, or bushiness, of a node is defined as the number of connections it has to other nodes in the graph. So, a highly linked node has a large overlapping vocabulary with several paragraphs, representing an important subtopic of the text. For this reason, it is a candidate for inclusion in the summary. Selection of highly connected nodes is done until the summary compression rate is satisfied in the Bushy Path. This way, the coverage of the main subtopics of the text is very likely to be good. However, the summary may be non-coherent, since relationships between every two nodes are not properly tackled. To overcome this, instead of simply selecting the most connected nodes, the Depth-first Path starts with some important node (usually the one weighted the highest) and continues the selection with the nodes (i) that are connected to the previous selected one and (ii) that come after it in the text, also considering selecting the most connected one among these, trying to avoid sudden subtopic changes. This procedure is followed until the summary is fully built. Its advantage over the Bushy Path is that more legible summaries may be built due to choosing sequential paragraphs. However, subtopic coverage is not guaranteed. The Segmented Bushy Path aims at overcoming the bottlenecks introduced by the other two methods. It tackles the subtopic representation problem by first segmenting the graph in portions that may correspond to the subtopics of the text. Then, it reproduces the Bushy Path method in each subgraph. This is done by selecting the most important paragraphs within a subtopic, and finally, uniting them with transitional paragraphs, which are chronologically prior to each first paragraph of each subtopic. It is guaranteed that at least one paragraph of each subtopic will be selected to compose the summary. In their evaluation, the authors showed that the methods produce good results for a corpus of encyclopedic texts, with the Bushy Path being the best one.

The first two methods adapted from [3] (Bushy Path and Depth-first Path) were already evaluated for MDS of texts written in Portuguese [12]. Using such methods for MDS, as we discuss later, implies in dealing with the multi-document phenomena, mainly with redundancy. The results, explained in more details in Section 4, are very promising, being close to the state of the art summarizers. It is important to note that, despite the use of paragraphs in [3] as the information unit to be selected, we chose to select sentences because of their more refined granularity, which allows for the construction of more informative summaries, as most

of the works in the area usually do nowadays. The system for Portuguese that incorporates Salton et al. methods is referred by RSumm.

Antiqueira et al. [17, 18] use complex networks to model texts for single document summarization. In their networks/graphs, each sentence is represented as a node and links are established among sentences that share at least one noun. Once the network is built, sentence ranking is performed by using graph and complex network measures, as degree, clustering coefficient and shortest path, and the best ranked sentences are selected to compose the summary. Using some of the measures, such method was also adapted for MDS summarization for the Portuguese language [20, 12], producing good results. Such system was named RC-Summ.

Castro Jorge and Pardo [8], in a deep approach, model several texts as just one graph, with nodes representing sentences and links representing discourse relations among the sentences. Discourse relations are based on the ones predicted by the Cross-document Structure Theory [21]. Such relations pinpoint similar and different sentence content, as well as different writing styles and decisions among the texts. For sentence selection, sentences that have more relations/links to others are preferred. This method is the one used by CSTSumm, cited in the introductory section. Cardoso [22] improves this method, incorporating discourse relations that happen for each single document (following the Rhetorical Structure Theory [23]) and considering topic/subtopic distribution in the source texts, using the subtopic delimitation method that we report in this paper. Following the original work, we refer to this method by the RCT-4 acronym (which indicates that it was the 4th method variation investigated by Cardoso, using RST, CST and Topics - each word contributes with the first letter to the acronym).

It is also relevant to cite two more summarization approaches, that are not directly based on graphs, but that were also tested for Portuguese. MEAD, proposed by Radev et al. [24], is a very popular summarization system. The tool incorporates multiple strategies for selecting sentences for summarization, namely: 1) the position of sentences in their documents; 2) lexical distance of sentences in relation to the centroid, i.e., the central sentence of the texts; 3) the longest common subsequences among the sentences; and 4) presence of keywords in the sentences. It has been widely used in the area and produces good results. It was tested for Portuguese in [12]. To the best of our knowledge, GistSumm [7] was the first MDS system for Portuguese. It consists in a very simple approach, using word frequency to compute the most important sentence of the source texts and, then, using lexical similarity among this sentence and the other ones, it selects the best ranked ones to compose the final summary. It is considered a baseline system in the area.

The above works are some of the main ones for MDS in Portuguese that have been evaluated on the same corpus (which is a benchmark for MDS in Portuguese) with the same evaluation metrics and setup. For these reasons, they are the ones that we compare our ap-

proach to, as we will show in Section 4. It is important to say that there are other few initiatives in MDS for Portuguese (e.g., the work of Silveira and Branco [25], that introduces the SIMBA system, which applies clustering and text simplification for summarization), but that use different evaluation setups and make any direct comparison unfair.

2.2 Subtopic Segmentation and Clustering

There are several approaches for subtopic segmentation, for written and spoken language, using different features and techniques. We focus here on some of the main and most used ones for written language, since this is the case of this paper.

One well-known and heavily used approach for subtopic segmentation is TextTiling [2], which is based on lexical cohesion. In its strategy, it is assumed that a set of lexical items is used during the development of a subtopic in a text, and, when that subtopic changes, a significant proportion of vocabulary also changes. For identifying major subtopic shifts, adjacent text passages of a pre-defined size (blocks) are compared for overall similarity. The more words these blocks have in common, the higher the chance that they address the same subtopic. Subtopic boundaries are established in points in the text that show representative lexical gaps.

Choi [26] developed the algorithm called C99, also based on lexical cohesion. Starting from preprocessed sentences, C99 initially calculates the similarity between each pair of sentences and produces a similarity matrix. From the matrix, it produces a rank-similarity: the more similar the sentences are with their neighbors, the higher the score will be. The lower ranks in the classification matrix indicate subtopic boundaries.

Riedl and Biemann [27], based on TextTiling, proposed the TopicTiling algorithm that segments documents using the Latent Dirichlet Allocation (LDA) topic model [28]. The documents that are to be segmented have first to be annotated with topic IDs, obtained by the LDA inference method. The topic model must be trained on documents similar in content to the test documents. The IDs are used to calculate the cosine similarity between two adjacent sentence blocks, represented as two vectors, containing the frequency of each topic ID. Values close to 0 indicate marginal relatedness between two adjacent blocks, whereas values close to 1 denote connectivity.

Du et al. [29] presented a hierarchical Bayesian model for unsupervised topic segmentation. The model takes advantage of the high modeling accuracy of structured topic models to produce a topic segmentation based on the distribution of latent topics. The model consists of two steps: modeling topic boundary and modeling topic structure.

Hovy and Lin [30] have used various complementary techniques, including those based on text structure, cue words and high frequency indicative phrases for subtopic identi-

fication in a summarization system. They argue that discourse structure might help subtopic identification. Following in this line, Cardoso et al. [31] showed that in fact discourse structure mirrors the subtopic segmentation.

In this paper, we used an adaptation of TextTiling [31] for the characteristics of the texts that we used, which are news texts written in Portuguese (different from the original proposal of TextTiling, which was for expository texts in English). As the authors of its adaptation show, TextTiling is among the best methods for subtopic segmentation of news texts, being robust and scalable, which are goals that we follow in this work. The performance of the adapted version was 77% precision and 40% recall.

Since the subtopic segmentation technique applied by TextTiling is linearly made, i.e., on only one document by time, the following problem may happen: if two subtopics from different texts are found, there is no guarantee that they correspond to distinct information; this may even happen inside the same document if some subtopic is interrupted by another one and then resumed later. Thus, in order to identify similar subtopics within and across documents, we need to cluster the previously segmented subtopics.

Clustering is a concept that arises naturally in many fields, where there are heterogeneous set of objects. It is natural to search for such methods to group/cluster objects based on some measure of similarity. For example, to set the distance between objects, it may be considered that the closer they are, the more similar they are. Thus, clustering is centered around an intuitive, but vague goal: given a set of objects, one may partition them into a collection of clusters in which objects in the same group are close, while objects in different groups are distant from each other.

There are many works in the area. For Portuguese, there is a tool named SiSPI [32], which performs clustering on the basis of the Single-pass algorithm [33] for clustering similar sentences. In this paper, we have simply adopted this solution, adapting it for clustering subtopics with some varied similarity measures. It appears to be a sufficient and suitable solution, allowing us to focus on the summarization method.

In the context of the discovery of related subtopics, the Single-pass algorithm requires a single sequential pass over the set of subtopics to be clustered. It is an incremental clustering algorithm (groups are incrementally created by analyzing all other previously created groups). Figure 4 shows the Single-pass algorithm, already adapted to the case of subtopics.

Input: A set D = <d_1, ..., d_n> with *n* documents, where each d_i = <s_1, ..., s_m> has *m* subtopics, for *n* and $m \geq 1$
Output: A set C = <c_1, ..., c_x> with *x* clusters, where each c_i = <s_1, ..., s_y> contains *y* subtopics, for $y \geq 1$

Step 1: Set the initial set *C* of clusters to be empty
Step 2: Select a subtopic s_i of a document d_i following a given order
    If *C* is empty
        Then add the first cluster to *C* by inserting a single element s_i
    Else compare s_i (treated as a new cluster with only one element) with all clusters in *C*
        If the similarity between s_i and any cluster *C* is above a predetermined threshold
            Then place s_i within the closest cluster in *C*
        Else add the new cluster to *C*
Step 3: Repeat the step above until all the subtopics of all documents are processed

Figure 4. Single-pass algorithm

Initially, the algorithm creates the first group by selecting the first subtopic of a collection of documents to be clustered. The algorithm iteratively decides whether a newly selected subtopic should be placed in a group already created or in a new one. This decision is made according to the specified similarity function, i.e., a predetermined similarity threshold. In this work, we use the cosine measure, a measure of lexical similarity [34], and also the occurrence of discourse relations among the subtopics (which we explain later in this paper). The higher the similarity value between two subtopics, the more similar they are. The threshold is chosen based on the calculation of the average similarity among all groups. We tested other similarity thresholds as well, and the average similarity produced good results and showed to be adaptable to different text types and genres.

2.3 Cross-document Structure Theory

The Cross-document Structure Theory (CST) [21] is used to describe discourse connections among topically related texts in any domain. The author proposed 24 CST relations, however, in this study we consider only 14 relations found in the corpus we use in our evaluation (see Section 4). Such relations are organized in a typology defined in [35][36]. This typology is shown in Figure 5.

The typology classifies CST relationships into two major groups: content and presentation/form. The content category refers to relations that indicate similarities and differences among contents in the texts. This category is divided into three subcategories: redundancy, complement, and contradiction. Redundancy includes relations that express a total or partial similarity among sentences. Complement relations link textual segments that elaborate, give continuity or background to some other information. The last subcategory for the content category only includes Contradiction. In the form category, all the relations that deal with

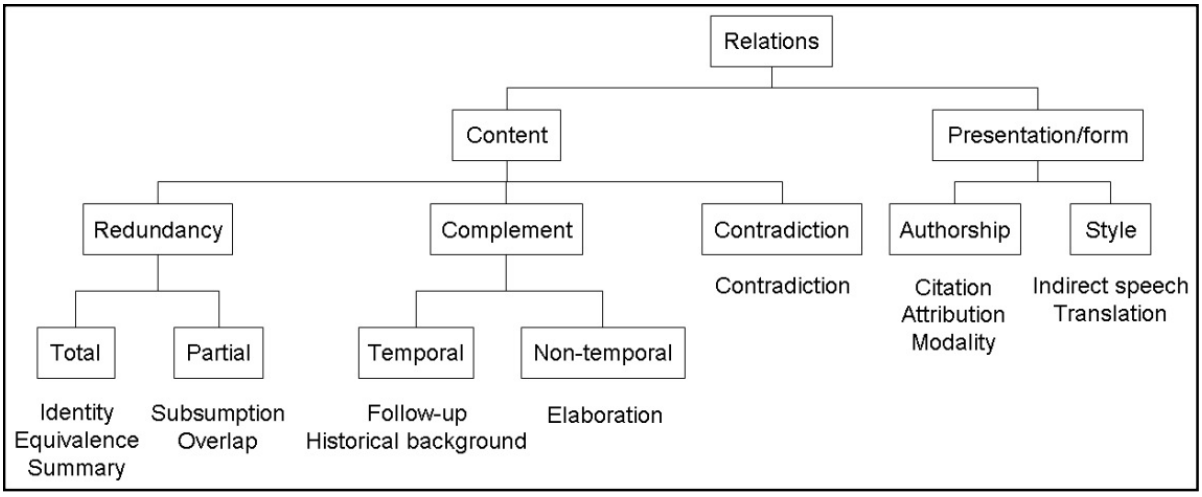


Figure 5. Typology of CST relations

superficial aspects of information are included.

Based on the meaning of the relationships defined in [35, 36], these may assist in the clustering phase, in order to obtain a better selection of similar subtopics. This is due to the fact that it was found in previous work that the discourse relations are closely related to the lexical similarity of textual segments, i.e., the closer such segments are, the more CST relationships they have with each other and, therefore, the greater the chance is that the segments belong to the same subtopic. Therefore, this model was used to investigate the issue of the relationship among subtopics.

The CST annotation of texts may be manually or automatically done. Manual annotation requires trained humans, making the process expensive and very time consuming. The automatic annotation, in turn, is performed by discourse parsers, which are softwares that automatically detect relationships among segments of texts and build graphs with the resulting annotation. For Portuguese, it is available a discourse parser for CST called CSTParser [35, 36], with a general accuracy of 68%. In this paper, we have used the manual annotation already available in our corpus, but, if one desires, the method may be scalable to new texts by using the available discourse parser.

In what follows, using the concepts and methods introduced before, we report our investigation and adaptation of the Salton et al. method under focus for MDS in Portuguese.

3 The Summarization Algorithm

The multi-document summarization method that we investigate in this paper - the Segmented Bushy Path - may be organized in a few steps. Firstly, the algorithm preprocesses the source texts and computes the lexical similarity among their sentences to build the map/graph. Then, it divides each text into subtopics using TextTiling (the adapted version to Portuguese news texts [31]). Once the texts are segmented, the next step is to identify and cluster common subtopics within and across the documents. With the resulting clusters, the relationship map is finally complete. The Segmented Bushy Path method may then be used to select the relevant information for the summary, performing the content selection. While the important information is being selected, the redundancy treatment is applied. In this way, it is guaranteed that the final summary does not have repeated information.

As mentioned before, the graph traversing was proposed by Salton et al.[3]. It is important to note that (1) we focus our investigation on the Segmented Bushy Path, due to the completion of the other two strategies in a previous effort [12], and (2) the chosen method was originally developed for the single document scenario, therefore, adaptations were made, in order that relevant information could be identified not only in one, but in several documents. We detail the main parts of the summarization method in the next subsections.

3.1 Preprocessing and Graph Building

We often encounter similar words in the source texts, but in different forms, e.g., house and houses. Therefore, a treatment for these types of words, which may be done by normalization (either by lemmatization or stemming [37]), is required. This treatment aims to standardize the words of the sentences and make the necessary computation more meaningful. As in other approaches, the stopwords are also eliminated using a pre-defined list for Portuguese language. Their removal is essential for keeping only the relevant words that carry the main content.

The preprocessed sentences in the texts are then used to build a graph, where the vertices are the sentences and links have numeric values that indicate how lexically close the sentences are (using the cosine measure).

Figure 6 shows an example of the performed preprocessing steps for two sentences and the computation of the cosine measure for them.

3.2 Subtopic Segmentation and Clustering

After the preprocessing, the texts are segmented in subtopics using TextTiling, as illustrated in Figure 7. The horizontal lines indicate where to hypothetically segment each

Considering two sentences of different documents that deal with the same topic, we have:

Sentence 1 (S1): All the houses were destroyed.
Sentence 2 (S2): The destruction spared no house.

Step 1 → Stopwords removal
S1: houses destroyed
S2: destruction spared house
Step 2 → Stemming of words
S1: hous destroy
S2: destruct spare hous
Step 3 → Application of the cosine measure
Step 3.1 → Binding of all distinct words
hous destroy destruct spare
Step 3.2 → Values assigned to each word according to the number of occurrences in each sentence
S1: 1 1 0 0
S2: 1 0 1 1
Step 3.3 → Finally, direct application of the cosine equation

$$\cos(S1, S2) = \frac{S1 \,.\, S2}{|S1|.\,|S2|} = 0.4082$$

Figure 6. Preprocessing step and similarity computation

document.

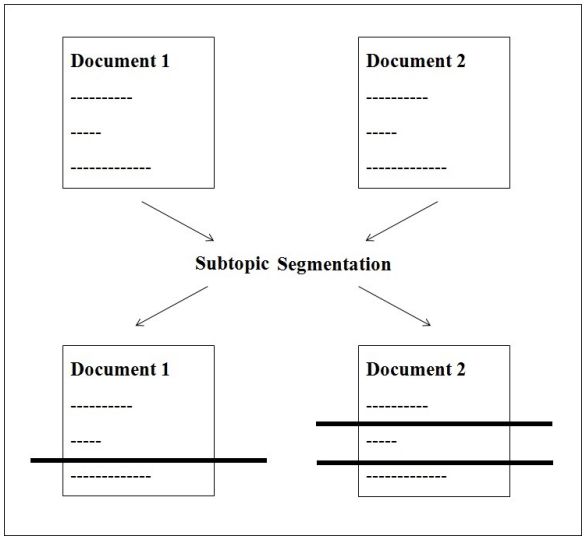


Figure 7. Subtopic segmentation by TextTiling

Since this technique was linearly applied for each document, as mentioned earlier, there is no correlation of the subtopics. Thus, it is necessary to perform subtopic clustering. For this, we have used the Single-pass algorithm [33] cited before. Figure 8 illustrates a possible (hypothetical) result for the clustering.

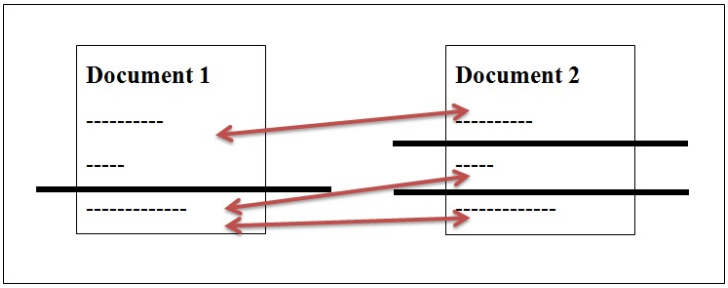


Figure 8. Clustering of subtopics

At this stage of subtopic correlation, 4 strategies for finding similar subtopics were considered: (1) Keywords - the most frequent words are discovered; then the cosine similarity is applied between the subtopics by analyzing only such words; (2) Subtopic similarity - all the words in the pair of subtopics are analyzed to determine whether they belong to the same cluster, using the cosine measure; (3) Unweighted CST - use of the number of CST relations among subtopics to investigate their correlation, i.e., the greater the number of connections between two subtopics, the greater the chance is of the subtopics being correlated; and (4) Weighted CST - use of numeric values for each CST relation between subtopics. These numerical values, in the range from 0 to 1, correspond to the level of similarity between each pair of subtopics. An Identity relation, for example, corresponds to the value 1 of similarity (since both segments are the same). For the Overlap relation, there is a value of 0.5 because of its aspect of partial redundancy: there may be a lot of information in common between sentences, but there may be also little redundancy. Thus, the greater the sum of the values of CST relationships between a pair of subtopics, the greater the chance is of grouping them.

As an illustration, Figure 9 shows the correlation of subtopics of two documents, after applying one of the above similarity techniques. It is a representation of subtopics in Figure 1. Sentences are indicated by *s* and subtopics by *sb*. Since sentences 1-2 of Document 1 and sentences 1-2 of Document 2 describe the same subtopic (Maradona's relapse), they must be grouped in a single cluster, identified by *sb1*. Sentences 3 and 5 of Document 1 are connected with sentences 3-5 of Document 2, forming subtopic *sb2*, since they describe details about Maradona's hepatitis. On the other hand, sentences 6-8 of Document 2 are not connected with any sentence of Document 1; in this case, they form a subtopic, identified by *sb4*. The same happens to sentence 4 of document 1. After the clustering, one may see that the set of texts has 4 subtopics.

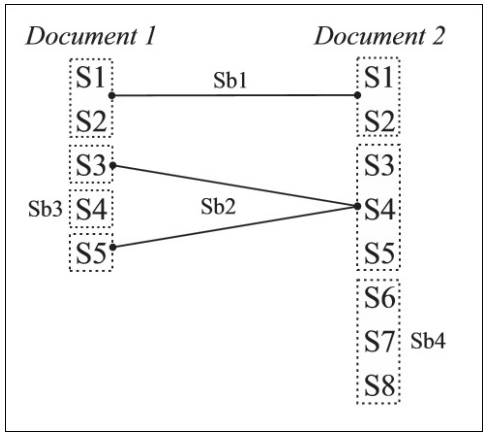


Figure 9. Correlation of subtopics

3.3 Content Selection for the Summary

After preprocessing, graph building, and subtopic segmentation and clustering, we have the content selection step, in which sentences are selected to compose the summary.

For a better overview of the method, Figure 10 shows the modeling of the source texts in a graph with the subtopic segmentation and clustering already performed. It is important to notice a few things about Figure 10: 1) the Si-Dj format indicates the location of a sentence (for instance, S1-D2 refers to the first sentence of the second document); 2) the lines between sentences must also indicate the cosine similarity they have; and 3) the strings above the segments indicate the keywords of the subtopics that they belong to (for illustrative purposes only, since our method do not need such topic signature words).

Having the graph, the application of Segmented Bushy Path is performed. As mentioned before, this path builds the final summary by selecting the most important sentences in each subtopic. For this, we choose the most connected sentence within a particular subtopic. For each new subtopic, it is also important to include a transitional sentence. This transitional sentence is always picked in a way that the chronology is maintained, i.e., the transitional sentence must come before the sentence of the next subtopic, being the one with the highest lexical similarity with the sentence of this new subtopic. This process of selecting sentences for each subtopic and then selecting transitional sentences occurs until the desired compression rate is reached.

It is noticeable that, using this method, many edges may possibly indicate a very high degree of similarity among the sentences (due to redundancy among texts). Therefore, it is necessary to calculate the limit of redundancy that two sentences may have with each other, establishing a threshold that enable to prune redundant sentences that might eventually be in-

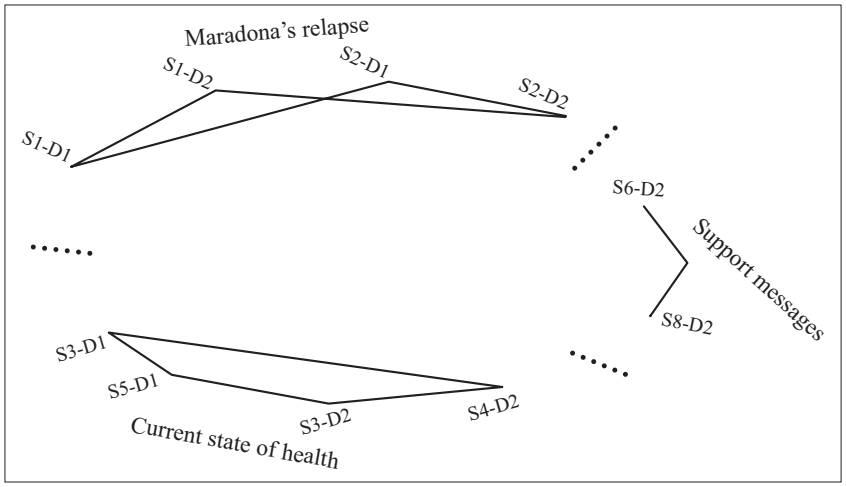


Figure 10. Relationship map

cluded in the summary. In other words, if a node (sentence) has a greater value of similarity (relative to the nodes that were already selected for the summary) than the established threshold, the sentence is not taken to the summary, as it is considered redundant. In this work, the threshold is computed as the average of the similarity values between every 2 sentences of every pair of documents. This way of establishing the threshold (instead of using a fixed value for any group of texts) allows the summarization method to be more generic and applicable to different types and collections of documents. We have also tested other threshold values and have empirically checked that the adopted strategy is a good solution.

Finally, the construction of the final summary should take into account its size, and, for this, we use a given compression rate (mentioned earlier), which limits the amount of information (counted in number of words, as usually done in the works in the area) that the summary will contain. We may notice that, just by applying the compression rate, relevant sentences may be left out of the summary in favor of transitional sentences. We opted to keep such an approach to, once again, preserve the summary coherence.

4 Experiments and Evaluation Results

In our experiments, we used the CSTNews corpus [38] for evaluating the MDS method ans also the subtopic segmentation and clustering steps. The CSTNews corpus is composed of 50 groups of news articles written in Brazilian Portuguese, collected from several sections (Politics, Sports, World, Daily News, Money, and Science) of mainstream online news agencies (Folha de São Paulo, Estadão, O Globo, Jornal do Brasil, and Gazeta do Povo). Specifically, each group contains 2 or 3 source texts on the same topic from different agencies (in a

total of 140 documents) and a multi-document human abstract for each group.

The corpus is a benchmark for researches on MDS for Portuguese and has been used in several studies (e.g., in all the MDS researches cited in the related work section). It is freely available and has several linguistic annotation layers that were manually produced in a systematic and reliable way. For this paper, we have special interest on the following: all the texts in the corpus were manually annotated with subtopic boundaries and their clustering, with satisfactory annotation agreement values (more details may be found in [39]); the texts present their CST manual annotation with satisfactory agreement values (as described in [38, 40]), which we tested in the context of subtopic clustering. For such reasons, we haved adopted this corpus instead of others frequently used in the area (as the corpora of the Document Understand Conferences).

The size of the human summaries in the corpus corresponds to 30% of the size of the longest text in each group (considering that the size is given in terms of number of words), resulting in a compression rate of 70%, therefore. This is the compression rate we use for producing the automatic summaries, in order to the comparison with the human summaries to be fair.

As we adopted a ready-to-use solution for performing subtopic segmentation (TextTiling), we start the evaluation by the clustering techniques we employed over the automatically identified subtopics. The quality of the clustering method may be assessed by external measures of quality that indicate how close the automatically produced clusters are in relation to the reference clusters (i.e., the clusters produced by humans in the corpus). For this evaluation, measures of precision (see Eq. 1), coverage (see Eq. 2) and f-measure (see Eq. 3) [41, 42] were used. Precision indicates the proportion of correct segments there are inside each cluster; coverage shows the proportion of correct segments there are in each cluster in relation to what was predicted in the reference clusters; f-measure is a unique performance measure, combining precision and coverage values.

$$P(k_i, c_j) = \frac{n_{ij}}{|c_j|} \tag{1}$$

$$C(k_i, c_j) = \frac{n_{ij}}{|k_j|} \tag{2}$$

$$F(k_i, c_j) = \frac{2 * C(k_i, c_j) * P(k_i, c_j)}{C(k_i, c_j) + P(k_i, c_j)} \tag{3}$$

In Equations 1-3, k_i indicates each reference cluster; c_j indicates each one of the clusters that are automatically formed; and n_{ij} refers to the number of segments of the cluster k_i that are present in c_j. The f-measure for each cluster of the entire data set is based on

the automatic cluster that best describes each reference cluster. Thus, the overall value of f-measure may be denoted by Equation 4.

$$Overall\ F - Measure = \sum_{k_i \in K} \frac{|k_i|\ max\ c_j \in C\{F(k_i, c_j)\}}{N} \quad (4)$$

In this formula, N refers to the total number of segments to be grouped; K is the set of reference clusters; and C is the set of automatic clusters.

The four clustering methods (Keywords, Subtopical Similarity, Unweighted CST and Weighted CST) were evaluated against the reference clusters and the results are presented on Table 1, using the Equations 1, 2, 3 and 4.

Table 1. Clustering results over automatic subtopic segmentation

Clustering Technique	Precision	Recall	F-Measure
Keywords	0.7265	0.6374	0.6790
Subtopical Similarity	0.5823	0.4165	0.4856
Unweighted CST	0.5514	0.3829	0.4519
Weighted CST	0.5490	0.3777	0.4475

From Table 1, we may conclude that the best technique, taking into account the automatic method of subtopic segmentation (TextTiling), was the one that simply considered the most frequent words in each subtopic for computing the cosine measure: the Keywords clustering method. This possibly happens because the technique considers only the most relevant words of a subtopic, so clustering is more accurate. It is also interesting to notice that all techniques had better precision than recall.

We went one step further and also evaluated the impact of using the reference (manual) segmentation of the corpus instead of the one automatically performed by TextTiling. The clustering results over the reference segmentation are shown in Table 2. As it may be seen, the values surpass those achieved by the automatic subtopic segmentation. The overrun was expected since we deal this time with better data. In addition, it may be noticed that the values obtained by Keywords are the highest among all the methods.

Table 3 presents how much better the clustering results achieved with the reference segmentation are in relation to the ones obtained with the automatic segmentation. For instance, one may see that the f-measure for the Keywords clustering over the reference segmentation is 6.44% better than the f-measure for the same clustering method over the automatic segmentation. In general, we may conclude that the results for the automatic methods are not far from those obtained for reference data, which favors their use in actual fully automatic systems.

Table 2. Clustering results over reference subtopic segmentation

Clustering Technique	Precision	Recall	F-Measure
Keywords	0.7586	0.6901	0.7227
Subtopical Similarity	0.6184	0.4300	0.5072
Unweighted CST	0.5680	0.4241	0.4850
Weighted CST	0.5324	0.4352	0.4789

Table 3. Improvement of clustering results - reference over automatic segmentation

Clustering Technique	Precision	Recall	F-Measure
Keywords	0.0442	0.0827	0.0644
Subtopical Similarity	0.0620	0.0324	0.0445
Unweighted CST	0.0301	0.1076	0.0732
Weighted CST	0.0302	0.1522	0.0702

The last and most important step of the evaluation phase corresponds to the analysis of the automatic summaries over the manual ones, which relied on the use of ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [43], which is a suite of metrics for this purpose.

Basically, ROUGE was created to enable a direct comparison between an automatically generated summary and its human references. ROUGE calculates a score among 0 and 1 based on sets of words (e.g., the n-grams that may vary from 1 to 4) in common between human summaries and automatically generated summaries, producing precision, recall/coverage, and f-measure values. As already mentioned before, precision indicates the proportion of reference n-grams in the automatic summary; recall indicates the proportion of reference n-grams in the automatic summary in relation to the reference summary; f-measure is a unique measure of performance, combining precision and recall.

It is known that ROUGE evaluates the informativeness of a summary, i.e., the amount of information the summary contains, and its scores have been shown to correlate well with human judgment. As it may be automatically and fastly computed, and achieves reliable results, it has become a mandatory evaluation metric and almost all the works in the area adopt it for evaluation. In this work, following what most of the works do, we consider ROUGE evaluation for n-grams of size 1 only (which we refer to by ROUGE-1), which is considered enough for distinguishing summary informativeness.

We initially used ROUGE to evaluate the versions of Segmented Bushy Path method that use the 4 clustering techniques previously tested, indicated by the numbers that follow the method name (1: clustering with all the words in the subtopics; 2: clustering with keywords only; 3: clustering using only the number of CST relations; 4: clustering using the weights of

the CST relations) and a version that uses the manually produced (reference) clusters in the corpus. We also included in the evaluation a random baseline, that randomly select sentences to compose the summary, for comparison purposes only. Table 4 shows the obtained results.

Table 4. ROUGE-1 results for segmented bushy path variations

Methods	**Precision**	**Recall**	**F-Measure**
Segmented Bushy Path^{Manual}	0.5803	0.3918	0.4407
Segmented Bushy Path^{1}	0.5472	0.3517	0.4190
Segmented Bushy Path^{2}	0.5507	0.3297	0.4023
Segmented Bushy Path^{3}	0.6079	0.2802	0.3571
Segmented Bushy Path^{4}	0.6033	0.2879	0.3637
Baseline	0.3015	0.2900	0.2948

As expected, the version with the reference clusters achieved the best results. Interestingly, looking at f-measure values for the automatic clustering versions, one may see that the best clustering technique (using the keywords) did not result in the best version of the summarization method. The best summarization method was the one that used all the subtopic words for clustering. This is not a surprise and probably happens due to variation in the summary content and due to the known fact in the area that several good summaries exist for the same group of texts. One may also see that all the method variations outperformed the baseline method.

In Table 5, there is a comparison with other summarizers for the Portuguese language, computed over the same corpus and with the same evaluation setup. The methods are ordered in the table by the f-measure values. The results correspond to the RSumm summarizers [12], RCT-4 [22], CSTSumm [8], MEAD [24], GistSumm [7] and RCSumm [20], making it possible to know the quality of the summarization method investigated in this paper in relation to other summarizers to Portuguese. The cited summarizers are the ones that were already introduced in the related work section. In addition, we included in the comparison existing versions of GistSumm and MEAD that were enriched with CST information (which help improving sentence ranking for performing sentence selection to compose the summary), which produce better results than their original versions.

Overall, one may see that the results were very satisfactory, with the investigated method overcoming not only most of the summarizers, but also the other two paths proposed in [3]. Our method only lost for the RCT-4 method, which is a discourse-based method, and, therefore, very expensive and not easily scalable.

The lower recall of our method in relation to RCT-4 may be explained by the fact that the priority was to select at least one sentence of each subtopic, and a transition sentence for new subtopics. Thus, compression rate was often hit up after including some transition

Table 5. Comparison among systems for Portuguese

Methods	Precision	Recall	F-Measure
RCT-4	0.4520	0.4416	0.4445
Segmented Bushy Path[1]	0.5472	0.3517	0.4190
RCSumm	0.4218	0.4036	0.4102
Segmented Bushy Path[2]	0.5507	0.3297	0.4023
MEAD with CST	0.4257	0.3876	0.4018
RSumm (Bushy Path)	0.4089	0.3704	0.3871
CSTSumm	0.4472	0.3557	0.3864
RSumm (Depth-First Path)	0.3977	0.3630	0.3795
Segmented Bushy Path[4]	0.6033	0.2879	0.3637
MEAD	0.3691	0.3574	0.3616
GistSumm with CST	0.2800	0.5229	0.3583
GistSumm	0.3923	0.3343	0.3581
Segmented Bushy Path[3]	0.6079	0.2802	0.3571
Baseline	0.3015	0.2900	0.2948

sentences, leaving out other important information (sentences of other subtopics) from the summary.

Another interesting measure we tested (but do not show here) is ROUGE-2. Its values for f-measure showed a different rank, in which the Segmented Bushy Path[1] obtained a higher f-measure value than the RCT-4 system (0.3434 vs. 0.2615, respectively), outperforming all the systems. Therefore, it is possible to claim that our method is competitive with the state of the art system, but at a lower cost, since it is a superficial method.

We have run t-tests for pairs of summarizers for which it was important to show that the differences in results were significant. The following pairs of comparisons were carried out: Segmented Bushy Path[1] RCT-4, Segmented Bushy Path[2] RCT-4, Segmented Bushy Path[1] RCSumm, Segmented Bushy Path[2] RCSumm, and Segmented Bushy Path[1] Segmented Bushy Path[2]. The results indicated that there is statistical difference with 95% confidence.

As illustration, Figure 11 shows the automatic summary produced by the Segmented Bushy Path[1] method for the texts previously presented in the introductory section (Figure 1). It may be noticed that not only Maradona's relapse (first subtopic) was reported, but also Maradona's hepatitis (second subtopic). The third and forth subtopics (Current State of Heath and Support Messages, respectively) are not included due to the choice of the compression rate (70%).

Maradona had been discharged 11 days ago, but he was again hospitalized on Friday, and the medical reports did not specify what was wrong with the ex-player – Cahe ruled out ulcer or pancreatitis.
Cahe said that Maradona had not started to drink alcoholic beverages again, and that the causes of the relapse are being investigated.

Figure 11. Example of automatic summary produced by the investigated method

Undoubtedly, there are some issues regarding the completeness and coherence of the generated summary, e.g., Cahe is an entity that lacks a clear reference: a reader unfamiliar with the events reported in the source texts is not able to determine that Cahe is Marodona's doctor. Still, the Segmented Bushy Path [3] proves to be effective when subtopic segmentation and clustering techniques are performed together to obtain a summary capable of comprising the most relevant subtopics in a group of texts.

5 Google Chrome Extension

Given the need for a tool that handles the large amount of online information, mainly in the current multi-document scenario, in this work we developed an extension for Google Chrome browser, which summarizes the documents returned from a search with Google News website. The extension makes use of the following tools: 1) the Application Programming Interface of Google News (Google News API) for retrieval of documents; 2) NCleaner tool [44], trained for Portuguese, for removing the irrelevant content in the web pages (advertisements and links to other pages, for example); and 3) the summarization methods of Salton et al. [3].

Figure 12 shows the active extension with the search for the term "Manifestações São Paulo" (in English, "Protests São Paulo", regarding the public acts that happened in the city of São Paulo, in Brazil) in Google News website, followed by Figure 13, where the return of the search was summarized (in this case, the first eight most relevant texts were processed).

It may be noted that, in Figure 12, for the search results in Google News site, the "Sumarizar" (in English, "Summarize") button appears at the top right corner, giving the user the option to summarize the retrieved texts. It is important to explain that the system is currently customized for the Portuguese language, since the stoplist and stemmer are used for this language. However, this customization may be easily made to other languages too, once such resources are available.

Figure 13 shows a summary with a compression rate of 70% on the search conducted earlier. It may be noted that, on the left, there are the links related to the search, as well as links

in each sentence of the summary (in the central panel) that connects to their corresponding source texts. Notice too that the users may add more texts to the process as well as increase or decrease the compression rate.

Figure 12. Google Chrome extension

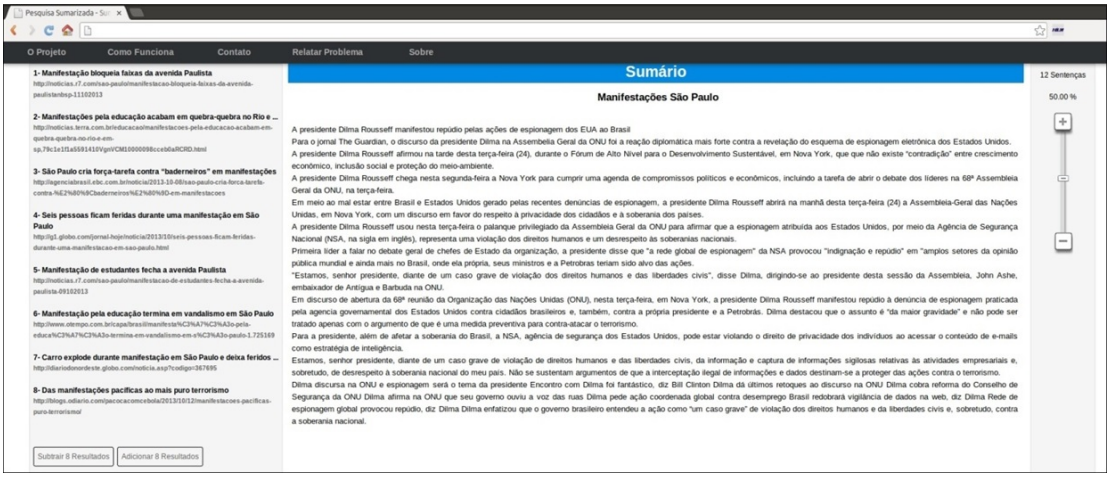


Figure 13. Generated summary in the online interface

6 Final Remarks

We presented a new multi-document summarizer with the Segmented Bushy Path method adapted from Salton et al. [3], which may summarize news articles. This summarizer not only uses Salton et al.'s technique of selecting the most relevant information, but also explores techniques of subtopic segmentation and clustering. The results show that the method is competitive with the best systems for Portuguese. In fact, since it is a superficial method, it is more scalable and capable of on-line integration than the current best available systems.

There is still room for improvements, as dealing with dangling anaphors and producing update summaries. To handle anaphors, or, in more general terms, co-references, is a very difficult task in Natural Language Processing area and is still a problem without robust solution. If we consider the Portuguese language, current systems still present severe limitations. Update summaries, on the other side, appear to be a more reachable and promising line for future work. Considering that possible users of our summarization system aims at browsing and summarizing news on the web, it makes sense to present to these users only the most relevant unknown information, filtering out the information that the users may already know. There are several update summarization techniques proposed in the literature that might be combined with our summarization strategy at relative low cost.

Finally, it is important to say that most of the tools and resources cited in this paper are publicly available. They may be found at the webpage of SUCINTO project [4], which aims at investigating and developing summarization strategies and the associated linguistic-computational resources, tools and applications, mainly for the Portuguese language.

7 Acknowledgments

The authors are grateful to FAPESP, CAPES, and CNPq for supporting this work.

Contribuição dos autores:

Os autores contribuíram igualmente para o desenvolvimento da pesquisa relatada e para a escrita do artigo.

[4] http://www.icmc.usp.br/pessoas/taspardo/sucinto

References

[1] I. Mani, Automatic Summarization, John Benjamins Publishing, 2001.

[2] M. A. Hearst, TextTiling: Segmenting text into multi-paragraph subtopic passages, Computational Linguistics 23 (1) (1997) 33–64.

[3] G. Salton, A. Singhal, M. Mitra, C. Buckley, Automatic Text Structuring and Summarization, Information Processing & Management 33 (2) (1997) 193–207.

[4] E. Hovy, Text Summarization, in: R. A. Lewin (Ed.), The Oxford Handbook of Computational Linguistics, Oxford University Press, The United States, 2009, pp. 583–598.

[5] L. Hennig, Topic-based multi-document summarization with Probabilistic Latent Semantic Analysis, in: the Recent Advances in Natural Language Processing, 2009, pp. 144–149.

[6] X. Xu, A new sub-topics clustering method based on semi-supervised learning, Journal of Computers 7 (10) (2012) 2471–2478.

[7] T. A. S. Pardo, GistSumm-GIST SUMMarizer: Extensões e novas funcionalidades, Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2005.

[8] M. L. R. Castro Jorge, T. A. S. Pardo, Experiments with CST-based multidocument summarization, in: the Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, Association for Computational Linguistics, 2010, pp. 74–82.

[9] X. Wan, An Exploration of Document Impact on Graph-based Multi-document Summarization, in: the Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 755–762.

[10] S. Harabagiu, F. Lacatusu, Topic themes for multi-document summarization, in: Proceedings of the 28th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2005, pp. 202–209.

[11] X. Wan, J. Yang, Multi-document summarization using cluster-based link analysis, in: the Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval, ACM, 2008, pp. 299–306.

[12] R. Ribaldo, A. T. Akabane, L. H. M. Rino, T. A. S. Pardo, Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information, in: the Proceedings of the 10th International Conference on Computational Processing of Portuguese (LNAI 7243), 2012, pp. 260–271.

[13] R. Mihalcea, D. Radev, Graph-based Natural Language Processing and Information Retrieval, Cambridge University Press, 2011.

[14] I. Mani, E. Bloedorn, Summarizing Similarities and Differences Among Related Documents, Information Retrieval 1 (1-2) (1999) 35–67.

[15] G. Erkan, D. R. Radev, LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, Journal of Artificial Intelligence Research (JAIR) 22 (1) (2004) 457–479.

[16] R. Mihalcea, P. Tarau, A Language Independent Algorithm for Single and Multiple Document Summarization, in: the Proceedings of the 2nd International Joint Conference on Natural Language Processing, 2005.

[17] L. Antiqueira, Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos, Ph.D. thesis, Universidade de São Paulo (2007).

[18] L. Antiqueira, O. N. Oliveira Jr, L. d. F. Costa, M. G. V. Nunes, A complex network approach to text summarization, Information Sciences 179 (5) (2009) 584–599.

[19] D. S. Leite, Um estudo comparativo de modelos baseados em estatísticas textuais, grafos e aprendizado de máquina para sumarização automática de textos em Português, Master's thesis, Departamento de Computação, Universidade Federal de São Carlos. São Carlos/SP, Brazil (2010).

[20] A. T. Akabane, T. A. S. Pardo, L. H. M. Rino, Sumarização multidocumento com base em métricas de redes complexas, Anais do 19o Simpósio Internacional de Iniciação Científica da Universidade de São Paulo-SIICUSP (2011) 1–1.

[21] D. R. Radev, A common theory of information fusion from multiple text sources step one: cross-document structure, in: the Proceedings of the 1st SIGdial workshop on Discourse and Dialogue, Association for Computational Linguistics, 2000, pp. 74–83.

[22] P. C. F. Cardoso, Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo, Ph.D. thesis, Universidade de São Paulo (2014).

[23] W. C. Mann, S. A. Thompson, Rhetorical Structure Theory: A theory of text organization, in: University of Southern California, Information Sciences Institute, no. ISI/RS-87-190, 1987.

[24] D. R. Radev, H. Jing, M. Budzikowska, Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies, in: the Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization, Association for Computational Linguistics, 2000, pp. 21–30.

[25] S. B. Silveira, A. Branco, Combining a double clustering approach with sentence simplification to produce highly informative multi-document summaries, in: IRI 2012: 14th International Conference on Artificial Intelligence, Las Vegas, USA, 2012.

[26] F. Y. Y. Choi, Advances in domain independent linear text segmentation, in: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, NAACL 2000, Association for Computational Linguistics, Stroudsburg, PA, USA, 2000, pp. 26–33.

[27] M. Riedl, C. Biemann, Topictiling: A text segmentation algorithm based on lda, in: Proceedings of ACL 2012 Student Research Workshop, ACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 37–42.

[28] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[29] L. Du, W. Buntine, M. Johnson, Topic segmentation with a structured topic model, in: Proceedings of NAACL-HLT, 2013, pp. 190–200.

[30] E. Hovy, C.-Y. Lin, Automated text summarization and the SUMMARIST system, in: the Proceedings of a workshop on held at Baltimore, Maryland, Association for Computational Linguistics, 1998, pp. 197–214.

[31] P. C. F. Cardoso, M. Taboada, T. A. S. Pardo, On the contribution of discourse to topic segmentation, in: the Proceedings of the 14th Annual SIGDial Meeting on Discourse and Dialogue, 2013, pp. 92–96.

[32] E. R. M. Seno, M. G. V. Nunes, Reconhecimento de Informações Comuns para a Fusão de Sentenças Comparáveis do Português, Linguamática 1 (1) (2009) 71–87.

[33] V. Rijsbergen, Information Retrieval, Butterworths, Massachusetts, 1979.

[34] G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[35] E. G. Maziero, T. A. S. Pardo, Automatic Identification of Multi-document Relations, the Proceedings of the PROPOR (2012) 1–8.

[36] E. G. Maziero, M. L. R. Castro Jorge, T. A. S. Pardo, Revisiting Cross-document Structure Theory for multi-document discourse parsing, Information Processing & Management 50 (2) (2014) 297–314.

[37] M. Porter, Snowball: A language for stemming algorithms. Available at: http://www.snowball.tartarus.org/texts/introduction.html.

[38] P. C. F. Cardoso, E. G. Maziero, M. L. R. Castro Jorge, E. R. M. Seno, A. Di Felippo, L. H. Rino, M. G. V. Nunes, T. A. S. Pardo, CSTNews - A Discourse-Annotated Corpus for Single and Multi-document Summarization of News Texts in Brazilian Portuguese, in: the Proceedings of the 3rd RST Brazilian Meeting, 2011, pp. 88–105.

[39] P. C. F. Cardoso, M. Taboada, T. A. S. Pardo, Subtopic annotation in a corpus of news texts: Steps towards automatic subtopic segmentation, in: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, 2013, pp. 49–58.

[40] P. Aleixo, T. A. S. Pardo, CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (Cross-document Structure Theory, Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2008.

[41] M. Steinbach, G. Karypis, V. Kumar, et al., A comparison of document clustering techniques, in: KDD workshop on text mining, Vol. 400, Boston, 2000, pp. 525–526.

[42] B. C. Fung, K. Wang, M. Ester, Hierarchical Document Clustering Using Frequent Itemsets., in: 3rd SIAM International Conference on Data Mining, Vol. 3, SIAM, 2003, pp. 59–70.

[43] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004, pp. 74–81.

[44] S. Evert, A lightweight and efficient tool for cleaning web pages, in: the Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008, Citeseer, 2008.

Avaliação para Trajetórias de Incisões Cirúrgicas com SVM

Assessment for Surgical Incision Trajectories with SVM

Ives Fernando M. S. de Moura [1]
Ronei Marcos Moraes [2]
Liliane dos Santos Machado [3]

Data de submissão: 26/04/2016, *Data de aceite: 11/05/2016*

Resumo: Simuladores de procedimentos cirúrgicos com Realidade Virtual vêm se tornando cada vez mais populares. Um componente importante deste tipo de *software* para a área médica é a capacidade de simular incisões. No contexto computacional, as incisões envolvem três etapas básicas: definição da geometria e topologia, detecção de colisão e deformação. Além disso, em simuladores voltados ao treinamento de estudantes, a capacidade de avaliar o desempenho do usuário é fundamental. Neste trabalho, dois experimentos são feitos, um com dados gerados aleatoriamente e outro com dados gerados a partir de interação do usuário com uma aplicação via mouse, para avaliar a qualidade do corte no que diz respeito à sua trajetória. Ambos os experimentos foram realizados utilizando o método *Support Vector Machine* (SVM). Resultados satisfatórios foram obtidos com a SVM e cinco *kernels*, com taxas de acerto superiores a 70%. Tem-se como principal conclusão, então, que a SVM é um método capaz de avaliar problemas relacionados à trajetória de incisão. Como contribuições do trabalho, têm-se a análise dos componentes da incisão, o levantamento de métodos de avaliação aplicáveis a este problema e o uso da SVM nos dois experimentos conduzidos, que mostrou a aplicabilidade deste método neste contexto.

***P*alavras-chave:** incisão, corte, cirurgia, avaliação, SVM

[1]Laboratório de Tecnologias para o Ensino Virtual e Estatística (LabTEVE), Universidade Federal da Paraíba (UFPB) - João Pessoa, Paraíba, Brasil. {ivesfmoura@gmail.com}
[2]Laboratório de Tecnologias para o Ensino Virtual e Estatística (LabTEVE), Universidade Federal da Paraíba (UFPB) - João Pessoa, Paraíba, Brasil. {ronei@de.ufpb.br}
[3]Laboratório de Tecnologias para o Ensino Virtual e Estatística (LabTEVE), Universidade Federal da Paraíba (UFPB) - João Pessoa, Paraíba, Brasil. {liliane@di.ufpb.br}

Abstract: Virtual Reality simulators for surgical procedures have been gaining popularity. An important component of this type of software for the area of health is the ability to simulate incisions. In the computational context, incisions involve three basic steps: definition of geometry and topology, collision detection and deformation. Furthermore, in simulators aimed for the training of students, the ability to assess user performance is fundamental. In this work, two experiments are conducted, one with randomly generated data e other with data generated through user interaction with an application using a mouse, to assess the quality of the incision with respect to its trajectory. Both experiments were conducted utilizing the Support Vector Machine (SVM) method. Satisfactory results were obtained with the SVM and five kernels, with success rates above 70%. The main conclusion is, therefore, that the SVM is a method capable of assessing problems related to incision trajectory. As contributions of the work, there is the analysis of the incision components, a survey of assessment methods applicable to this problem and the use of the SVM in both experiments, which has shown the applicability of this method in this context.

***K*eywords:** incision, cutting, surgery, assessment, SVM

1 Introdução

A realização de tarefas na área da saúde está condicionada, muitas vezes, ao uso de materiais que são caros ou de difícil acesso, podendo trazer, ainda, riscos à saúde de quem as estiver realizando. Estes fatores servem, então, como motivação para o uso de softwares capazes de realizar ou simular estas tarefas [1]. Programas de computador são empregados nesta área, por exemplo, no treinamento de atuais ou futuros profissionais que desejam adquirir conhecimento sobre procedimentos que podem ser cirúrgicos ou não [2].

No contexto de simuladores para procedimentos cirúrgicos, a possibilidade de realizar incisões (ou cortes) tem papel fundamental, já que é uma etapa presente em grande parte das atividades dessa natureza. A realização destes cortes nos simuladores traz, no entanto, uma série de desafios para o desenvolvimento dessas ferramentas, uma vez que para que se tenha um resultado fisicamente realista é necessário fazer uso de diversas ferramentas matemáticas, físicas e computacionais [3].

Mesmo com esta complexidade associada, os simuladores com Realidade Virtual (RV) vêm se tornando cada vez mais populares no treinamento de estudantes, já que além das vantagens citadas, eles promovem ainda a possibilidade de realizar o mesmo procedimento repetidas vezes, sem a necessidade de subtituir ou repor os materiais utilizados após cada uso. Além disso, métodos para acompanhamento e para a avaliação dos estudantes podem ser integrados a este *software*, podendo fornecer, então, *feedback* aos usuários, para que possam

observar seus erros e ter conhecimento de seu desempenho na realização do procedimento em questão [4].

No contexto de simuladores para procedimentos cirúrgicos, a possibilidade de realizar incisões (ou cortes) tem papel fundamental, já que esta é uma etapa presente em grande parte das atividades dessa natureza. O treinamento para realização de incisões se mostra importante, pois se trata de uma habilidade psicomotora com a qual os alunos têm contato apenas quando já o estão vivenciando em situação prática, o que pode vir a ocasionar insegurança e ansiedade aos alunos, bem como oferecer riscos às partes envolvidas [2]. Com um simulador é possível transmitir para os estudantes a noção do que acontece de fato em uma sala de cirurgia, preparando-os para a prática concreta, uma vez que eles já haverão experimentado situações similares.

Simuladores comerciais, como os listados em [5], que têm suporte a procedimentos com incisão e também avaliação de usuário já são utilizados em contextos hospitalares e universitários atualmente, mas eles possuem a desvantagem de terem alto custo associado. Simuladores com código aberto e sem fins comerciais também existem, mas este tipo de *software* não é de desenvolvimento simples, requerendo extensa pesquisa e equipamentos específicos, o que dificulta a disponibilização dos simuladores para uso livre e com baixo custo.

Será feita, neste trabalho, uma discussão sobre técnicas de incisão utilizadas em simuladores de procedimentos cirúrgicos com RV, mostrando seus requisitos básicos e seu funcionamento. São mostrados também exemplos de softwares deste gênero que já foram desenvolvidos ou continuam em desenvolvimento. Serão expostos, ainda, modelos de decisão que são empregados na avaliação do desempenho de estudantes ao utilizar simuladores deste tipo e as variáveis envolvidas neste processo, uma vez que este é um componente importante em softwares de treinamento. Isto é notável especialmente em se tratando de uma habilidade psicomotora relacionada à área de saúde, como é o caso das incisões, já que, geralmente, os alunos têm contato com este tipo de procedimento apenas quando já estão o vivenciando em situação prática, o que pode gerar ansiedade e insegurança, ocasionando erros e riscos [2]. Neste contexto de avaliação são feitos, por fim, dois experimentos utilizando o modelo *Support Vector Machine* (SVM) para avaliar dados sobre trajetórias de corte. A análise dos componentes da incisão, o levantamento de métodos de avaliação aplicáveis a este problema e o uso da SVM nos dois experimentos conduzidos se destacam como principais contribuições deste trabalho.

2 Corte virtual

Patnaik, Singla e Bansal [6] e outros autores destacam três características essenciais para as incisões: a acessibilidade, a extensibilidade e a segurança. Acessibilidade é relaci-

onada com a capacidade de se alcançar a estrutura anatômica que será explorada e realizar o procedimento necessário. A extensibilidade tem relação com a habilidade de aumentar a incisão em uma direção que facilite a operação, mas não danifique as funções do corpo. Finalmente, segurança neste contexto se relaciona a manutenção das funções corporais do paciente e o cuidado com a vizinhança do local da incisão. É importante, então, que sistemas computacionais que planejam lidar com a simulação de cortes sejam capazes de representar estes aspectos em seu mundo virtual.

A simulação de técnicas de incisão com RV para softwares voltados ao uso na área de saúde deve se preocupar em ser eficiente, robusta e realista, além dos aspectos de saúde já citados. Para isso, três componentes se destacam na implementação do chamado corte virtual: a detecção de colisão entre o intrumento de corte e o objeto a ser cortado, a simulação do corpo como um objeto deformável (com características de elasticidade) e, finalmente, a atualização da representação geométrica e topológica do modelo utilizado para comportar o corte [3].

Na maioria das vezes, os objetos tridimensionais vistos no computador ao utilizar um simulador são representados internamente como um conjunto de polígonos ou poliedros (triângulos ou tetraedros, em geral) chamados de primitivas. A conformação destes objetos geométricos no espaço, isto é, o posicionamento deles e sua vizinhança, forma a topologia do objeto. O corte é, então, uma operação topológica que rearranja as primitivas, subdivindo-as e movimentando-as para gerar uma lacuna no objeto. Os instrumentos de corte (bisturis, lâminas, serras, etc.) também serão representados no simulador como objetos compostos por primitivas. Para saber, então, quando realizar o corte, o sistema deve ser capaz de detectar quando e onde as primitivas do instrumento e do órgão que será cortado estão em contato. Esta tarefa recebe o nome de detecção de colisão [7]. Finalmente, ao tocar com o instrumento de corte no órgão e ao realizar o corte, é esperado que haja contração ou relaxamento dos tecidos envolvidos no processo, já que os tecidos do corpo humano apresentam elasticidade. À simulação destes fenômenos físicos, dá-se o nome de deformação

O procedimento de corte tem início a partir do momento em que há colisão entre o instrumento de corte e o órgão e é finalizado quando esta colisão deixa de ocorrer. À medida que o instrumento é movimentado, a modificação na topologia do órgão é feita e a deformação é aplicada, conforme mostrado no diagrama da Figura 1, elaborado a partir do exposto em [8] e [9]. Outros fenômenos podem ser acrescidos à simulação para aumentar seu realismo, como o uso dos chamados dispositivos hápticos, que fornecem retorno de força ao usuário, para que ele possa sentir de fato o que está tocando no mundo virtual; e a aplicação de técnicas de dinâmica de fluidos, para a simulação de sangramentos.

Do ponto de vista computacional, o corte é uma operação bastante custosa e, por isso, várias abordagens são utilizadas em sua implementação, variando a forma como os três componentes básicos aqui descritos são tratados. Uma discussão mais aprofundada sobre este

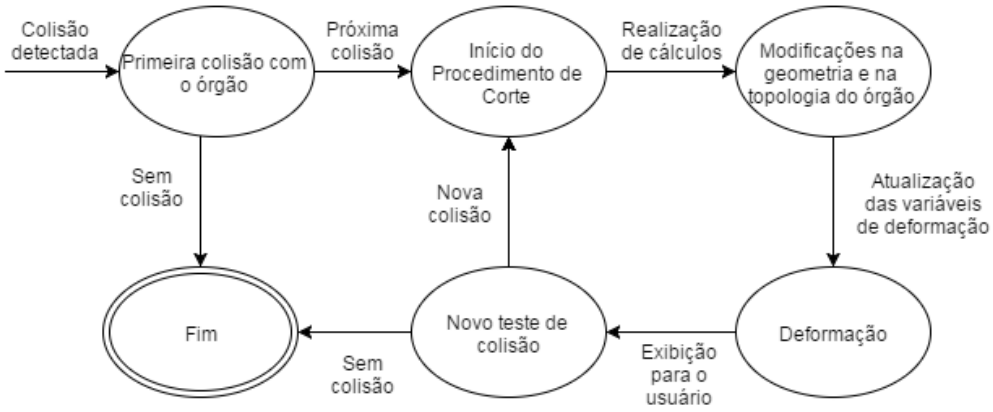


Figura 1. Diagrama do procedimento de corte

assunto pode ser encontrada em [10] e [8].

3 Simuladores de Procedimentos Cirúrgicos com RV

Existe, na literatura, uma série de exemplos de simuladores desenvolvidos para diversos procedimentos cirúrgicos diferentes. Serão citados aqui alguns simuladores cirúgicos com suporte a procedimento de incisão, destacando suas características em relação à forma como realizam o corte.

Com relação à representação geométrica e à topologia dos objetos tridimensionais do simulador, é convencional utilizar objetos compostos por triângulos ou tetraedros. Existem, no entanto, simuladores que fogem dessa convenção, utilizando outras representações para seus objetos. Um exemplo disto pode ser encontrado em [11], que representa seus objetos como nuvens de nós, computando o corte a partir do chamado critério de visibilidade, utilizando o método *level set*.

Já no que diz respeito à detecção de colisão, existe enorme variabilidade dentre os simuladores encontrados. Esta característica é esperada, no entanto, já que a detecção de colisão é completamente dependente do tipo de primitiva utilizado (tanto nos órgãos quanto nos intrumentos, que não necessariamente utilizam o mesmo tipo), existindo ainda diversas estratégias para a otimização desta tarefa, em vista de seu custo computacional. Uma discussão aprofundada sobre detecção de colisão pode ser encontrada em [12]. Como exemplos de diferentes abordagens para detecção de colisão, podemos citar [4], que se utilizou de método generalizado utilizando cilindros; [13], que se utiliza da distância do instrumento ao plano de corte e métodos analíticos; e [10], que fazem uso de uma abordagem baseada na topologia descrita por eles, chamada de Elementos Finitos Contínuos, que se utiliza de primitivas

hexaédricas.

Finalmente, no tocante à deformação, duas técnicas específicas são mais comumente encontradas: o modelo Massa-Mola, exemplificado em [13], e a técnica de Elementos Finitos, que pode ser encontrada em [14]. Há, no entanto, outras técnicas disponíveis, como visto em [15], que utilizam uma técnica de dinâmica baseada em posição; e em [16], que utilizam a chamada *Free-Form Deformation* (FFD).

Um grande número de simuladores com suporte a incisão existe na literatura, empregando diversas técnicas para alcançar seus objetivos. É interessante observar que, apesar de existir uma predominância de simulações para procedimentos laparoscópicos, ainda assim é possível encontrar simuladores que tratam de procedimentos que requerem incisões em maior escala ou que não comportam laparoscopia, como pode ser visto em [17] e [18], por exemplo.

4 Treinamento e Avaliação de Incisões com Simuladores

As possibilidades de reduzir custos e riscos, substituir materiais caros ou de difícil acesso e de repetir a execução de um determinado procedimento várias vezes sem a necessidade de utilizar novos materiais a cada vez ajudam a tornar o uso de simuladores para o treinamento de estudantes atrativo [19]. Outro aspecto interessante neste contexto é a possibilidade de promover a colaboração entre profissionais e estudantes de diferentes áreas durante o treinamento, através dos chamados Ambientes Virtuais Colaborativos [20].

A partir das interações dos usuários com o simulador é possível extrair diversas informações que podem ser utilizadas posteriormente para realizar a avaliação do desempenho do indivíduo [19]. No caso das incisões, o sistema pode coletar informação sobre posição do instrumento de corte, força aplicada, extensão do corte, dentre outras, que podem ser então comparadas com informações presentes na literatura da área ou obtidas através de conhecimento especialista. Essas informações, que neste contexto são chamadas de métricas, constituem uma limitação para os sistemas de avaliação, uma vez que elas dificilmente são bem-definidas e seguidas rigidamente no mundo real [21].

Esta característica de inexistência ou negligência de métricas introduz grande incerteza em relação aos valores medidos a partir da simulação, pois torna-se complexo definir o que é certo ou errado com precisão. Além disso, é possível tratar também de variáveis qualitativas, por exemplo, questões estéticas relacionadas ao corte. Vê-se, então, que além de não existirem métricas universalmente aceitas, os diferentes tipos de variáveis existentes no problema também influenciam na avaliação do usuário. É necessário, portanto, escolher um modelo de decisão que seja capaz de se adequar a essas restrições.

Paiva et al. [22] descrevem quatro modelos de decisão comumente utilizados em simuladores com RV. A Tabela 1 traz estas informações de maneira adaptada, incluindo o

Tabela 1. Diferentes modelos de decisão, suas descrições e tipos de variáveis suportados. Adaptada de [22] (p. 5)

Modelo de Decisão	Descrição
Lógica Clássica	Utilizada quando o problema possui estados com relação de causalidade bem definida
Lógica *Fuzzy*	Utilizada quando se possui dados com imprecisão ou incerteza
Sistemas Especialistas Baseados em Regras	Utilizados para modelar cenários definidos a partir de conhecimento de especialistas humanos
Árvore de Decisão	Utilizada para organizar o conhecimento a partir do ganho de informação associado aos atributos do problema
Support Vector Machine	Utilizada para agrupar dados em conjuntos a partir do uso de estruturas chamadas hiperplanos

modelo *Support Vector Machine*, o qual será discutido posteriormente neste trabalho. A avaliação do corte pode tratar de variáveis ordinais, intervalares ou nominais, dependendo do que for levado em consideração para classificar a incisão como boa ou não. Além disso, como existe imprecisão nas variáveis, elas podem ser modeladas como variáveis *fuzzy*, que tratam exatamente deste tipo de problema [23].

Quatro métricas são definidas em [24] para avaliar performance em treinamento de laparoscopia: o tempo para completar a tarefa, o comprimento do caminho percorrido pelo instrumento de corte, o volume da menor caixa que contém todas as amostras capturadas com o sensor na ponta do instrumento utilizado e o esforço de controle, que tem relação com as forças que o usuário gerou durante a simulação, as quais são calculadas analiticamente a partir dos dados obtidos. A partir destas ideias, é possível delinear métricas similares para a avaliação de tarefas de incisão.

Com a técnica de detecção de colisão utilizada, serão obtidas primitivas, as quais podem ser utilizadas então para construir o caminho do corte, sendo possível extrair, portanto, informações sobre comprimento e trajetória, por exemplo. Além disso, de acordo com o dispositivo que se utilizar para interagir com o sistema, será possível obter também dados de profundidade, força, aceleração, etc. Se a simulação for capaz de exibir e processar não apenas o órgão que será cortado, como também sua vizinhança, é possível avaliar aspectos de segurança no corte, no sentido de verificar se o usuário atingiu ou não regiões que não deveria. Podem-se incluir ainda outros aspectos, como os estéticos ou os de biossegurança, mas é importante perceber que a cada variável incluída na avaliação, o sistema se torna mais complexo.

Além da complexidade, tem-se o problema de definição de métricas, como já citado. Por mais que haja definições ou guias para a realização de incisões específicas, como pode ser visto em [6], os valores contidos na literatura da área não são necessariamente aceitos pela comunidade, podendo haver divergências entre as noções de corretude entre diferentes especialistas na área.

A SVM, utilizada em [24] como forma de avaliar a performance do usuário em procedimento laparoscópico com base nas quatro métricas definidas apresentou bons resultados. Devido à similaridade entre as métricas deste trabalho e as envolvidas na avaliação de incisões, a SVM se mostra uma candidata interessante para esta tarefa.

A SVM é um método proposto originalmente na década de 60, mas que evoluiu ao longo das décadas, sendo sua versão atualmente mais conhecida, referida como *soft margin*, descrita em [25].A ideia inicial da SVM é que dadas duas classes de dados, é possível construir um hiperplano que as separe, de forma que novos dados possam ser classificados de acordo com sua posição relativa ao hiperplano. Esta ideia pode ser então estendida para situações em que há mais de duas classes (conjuntos de hiperplanos) e, ainda, para casos em que os dados não são linearmente separáveis, através de um artifício chamado de *kernel*.

A versão linear da SVM procura escolher o melhor hiperplano de separação a partir do conceito de margens. As margens do hiperplano são as distâncias do plano aos dados mais próximos dele em cada classe. A SVM busca, então, maximizar essa distância mínima, fazendo o problema se tornar um clássico problema de otimização [26], representado pela Equação 1, onde m é o número de elementos do conjunto de treino, $cost_0$ e $cost_1$ são as funções de custo associadas à pertinência de uma amostra à classe ou não, $y^{(i)}$ é a variável de saída para a amostra i, $x^{(i)}$ é o vetor de variáveis de entrada para a amostra i, θ^T é o vetor de parâmetros transposto e o termo final é o chamado termo de regularização, que é utilizado para lidar com o problema de *overfitting* [27].

$$min_\theta C = \sum_{i=1}^{m}[y^{(i)}cost_1(\theta^T x^{(i)}) + (1 - y^{(i)})cost_0(\theta^T x^{(i)})] + \frac{1}{2}\sum_{i=1}^{n}\theta_j^2 \qquad (1)$$

Para tratar, por fim, de casos que não são linearmente separáveis, utilizam-se os *kernels*, que podem ser entendidos como métodos para alterar a dimensionalidade dos dados, mapeando-os para outro espaço vetorial, onde seja possível separá-los linearmente [28]. Concretamente, a hipótese da SVM (que computa a Equação 1 para decidir a que classe uma amostra pertence) pode ser reescrita como a soma do vetor de parâmetros θ multiplicado por um novo vetor de características f, como mostrado na Equação 2. Cada elemento do vetor de características pode ser entendido, então como uma função de similaridade (f_i) entre os valores de entrada (x) e as chamadas *landmarks* (l_i), que são pontos definidos no espaço do problema como representativos das classes, como mostrado na Equação 3. Estas funções de

similaridade são chamadas, finalmente, de *kernels* [29].

$$h_\theta(x) = \theta_0 + \theta_1 f_1 + \theta_2 f_2 + ... + \theta_n f_n \tag{2}$$

$$f_i = similaridade(x, l_i) \tag{3}$$

Uma série de conceitos matemáticos e estatísticos estão envolvidos na concepção das SVMs, trazendo várias vantagens consigo, como o suporte a dados não-linearmente separáveis e a alta dimensionalidade de dados. Este modelo é explorado de forma mais detalhada em [25].

5 Avaliação de trajetórias de corte utilizando SVMs

A partir da técnica de detecção de colisão utilizada, em conjunto com um dispositivo de interação, é possível obter uma lista com as primitivas intersectadas ao longo do movimento realizado. Estas primitivas, que podem ser pontos, triângulos, etc., podem ser então utilizadas para discretizar o caminho de corte. Define-se, então, a trajetória do corte como um conjunto de primitivas, no qual se destacam as primitivas inicial e final do corte, e todos os outros elementos do conjunto, quando conectados, formam um caminho coeso entre as duas extremidades. Coeso, neste contexto, significa que o caminho tem formato similar ao seu equivalente contínuo e ao que se esperaria no mundo real, não havendo desvios significativos.

Se for considerado um procedimento qualquer que exija uma incisão retilínea, o caminho ideal seria, claramente, uma reta que ligasse estes dois pontos, como visto na Figura 2. Em uma situação real, no entanto, dificilmente este caminho de corte será obtido, pois podem existir variações no posicionamento dos pontos inicial e final por parte de quem estiver realizando a incisão e, também, raramente será possível produzir uma incisão perfeitamente retilínea a mão livre. É preciso, então, especificar o que é considerado um bom ponto inicial (ou ponto final) ou não. Uma abordagem possível para isso é definir regiões em torno do ponto, conforme pode ser visto na Figura 2. A região mais interna conteria os pontos considerados bons, a do meio conteria os pontos considerados aceitáveis e a mais externa, os pontos considerados ruins. Pontos fora destas regiões seriam desconsiderados, pois não refletiriam um comportamento real, já que espera-se que um usuário, ao utilizar o simulador, esteja fazendo esforço para permanecer no caminho correto. Com esta primeira ideia, cortes bons seriam aqueles que ligam pontos em regiões boas, cortes aceitáveis são aqueles em que pelo menos um dos pontos está na região aceitável e cortes ruins seriam aqueles em que pelo menos um dos pontos está na região ruim.

O problema com esta abordagem é que ela não considera casos de fronteira, isto é, casos em que poucos pontos se localizam fora de determinada região, mas a maioria restante

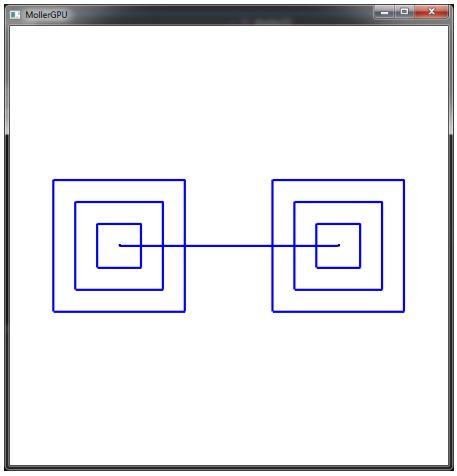


Figura 2. Exemplo de corte retilíneo ideal e regiões de classificação para os pontos inicial e final

está dentro dela acabam recebendo classificação indevida por causa da minoria que ficou de fora. Além disso, a noção de trajetória também acaba perdida, já que coloca-se ênfase apenas nos pontos incial e final, tornando os intermediários irrelevantes. Para corrigir este problema, deve-se estender as regiões para englobar não apenas os pontos extremos, mas todo o corte, criando-se então regiões de cortes bons, aceitáveis ou ruins. Além disso, deve-se considerar a que regiões os pontos coletados pertencem, atribuindo a categoria da maioria dos pontos ao corte. É importante observar, ainda, que esta ideia pode ser aplicada também a trajetórias com outros formatos, desde que se delimite corretamente as regiões boa, ruim e mediana e que se amostre uma quantidade suficiente de pontos.

De posse dessas duas ideias, pode-se construir dois experimentos, com o objetivo de avaliar se há a possibilidade de classificar cortes nas categorias definidas (Bom, Aceitável ou Médio e Ruim). As próximas seções deste artigo descrevem a metodologia utilizada e os resultados obtidos.

6 Metodologia

Com base nas duas ideias desenvolvidas anteriormente, dois experimentos foram realizados, com o objetivo de avaliar a viabilidade de utilizar as representações explicadas para a avaliação de cortes, no que diz respeito à trajetória deles.

Para o primeiro experimento, definiu-se que o corte seria retilíneo, partindo do ponto (10,10) para o ponto (20,10). Definiu-se, ainda, que a região considerada boa teria raio 1,0 ao redor dos pontos, a região aceitável teria raio 2,0 e a região ruim, 3,0. Foram gerados, então,

50 pontos aleatórios dentro de cada uma das regiões, tanto para o ponto inicial, quanto para o final. Após isso, 60 pares de pontos foram gerados para cada categoria e classificados de acordo com as regiões dos pontos, formando então o primeiro conjunto de dados, com 180 amostras. Para integrar o banco de dados, foi calculada a equação da reta para cada par de pontos, e pontos intermediários foram obtidos, gerando, finalmente, múltiplas retas. O banco de dados consiste então de 60 amostras para cada classe, e cada amostra representa um par de pontos aleatoriamente gerado dentro das regiões especificadas e 9 pontos intermediários entre eles.

No segundo experimento, o corte retilíneo foi mantido, mas a forma como os dados foram gerados foi diferente. Os dados agora passaram a ser capturados a partir da interação do usuário com a aplicação, via mouse, o que representaria a segunda ideia explorada na seção anterior, capturando os pontos intermediários diretamente a partir da interação do usuário. A aplicação capturava a posição do mouse no espaço da imagem (posição dos pixels) a partir do momento em que o usuário clicasse em algum ponto (sendo este seu ponto inicial) até que ele parasse de pressionar o botão (momento este em que seria registrado o ponto final). Como o volume de dados gerado dessa maneira é grande, foi feita posteriormente uma amostragem de 11 pontos por corte realizado, formando finalmente o segundo conjunto de dados. Este conjunto, assim como o primeiro, também contém 60 amostras de cada classe, tendo então um total de 180 amostras, mas agora os dados são classificados de acordo com a região a que a maioria dos pontos registrados durante a interação pertence, isto é, se todos os pontos estiverem na região considerada boa, a trajetória será considerada boa; se houver pontos apenas da região mediana ou tanto da mediana quanto da boa, o corte é considerado mediano e, finalmente, aquelas trajetórias que contiverem pontos da região ruim, sejam elem exclusivos ou em conjunto com pontos das outras regiões, são consideradas ruins. Todo o conjunto de dados desse experimento foi gerado a partir da interação do desenvolvedor do programa com a aplicação, estando este ciente dos parâmetros utilizados e da forma como deveria realizar o procedimento, evitando então a inclusão de casos que não refletem a realidade, como trajetórias em zigue-zague, por exemplo.

Para a criação de ambos os conjuntos de dados, foi feito um programa, utilizando a linguagem de programação C++. Na geração do primeiro conjunto, destaca-se o uso da biblioteca *random* da linguagem C++, em especial os componentes *random_device*, que busca gerar números aleatórios de forma não-determinística através de processos estocásticos, e a classe *uniform_real_distribution*, que utiliza o dispositivo para gerar números aleatórios com distribuição uniforme real. Já para gerar o segundo conjunto de dados, foi utilizada a biblioteca gráfica OpenGL e a biblioteca de utilitários GLUT, para que fosse possível fornecer a visualização dos pontos e das regiões de corte e gerenciar a interação do usuário com o ambiente, salvando os pontos obtidos com o movimento do mouse.

Construídos os conjuntos de dados, partiu-se para a avaliação deles com o uso de

um modelo de decisão específico, neste caso, as *Support Vector Machines* (SVMs). Para ter acesso a esse modelo, utilizou-se o software Weka (versão 3.6.12), que conta com duas implementações deste método, a libSVM e a SMO, cujas implementações das SVMs diferem pelos métodos numéricos que utilizam internamente. Os algoritmos implementados por ambos os métodos são explicados em [26] e [30], respectivamente.

Para avaliar os resultados obtidos ao fim de cada execução da SVM, utilizou-se o coeficiente Kappa, que é fornecido pelo Weka no relatório de cada execução. O coeficiente Kappa é uma medida de concordância, bastante utilizada na literatura de avaliação como forma de medir o desempenho em relação à quantidade de casos corretamente avaliados, tendo sido proposto por Cohen (1960). Este coeficiente assume valores entre 0 e 1 e, quanto mais próximo de um, melhor o modelo é capaz de descrever os dados, gerando avaliações mais precisas. Com o objetivo de encontrar que versão da SVM melhor se adaptaria a estes dados, 7 *kernels* diferentes foram utilizados com cada conjunto de dados, 3 com o método libSVM (linear, polinomial e RBF) e 4 com o SMO (polinomial, polinomial normalizado, Puk e RBF), sendo registrados os coeficientes Kappa e as taxas de acerto para cada execução, como mostrado nos resultados.

7 Resultados e Discussão

O modelo SVM foi empregado nos dois experimentos descritos anteriormente. Para cada experimento 7 *kernels* diferentes foram utilizados por meio do *software* Weka e suas implementações de SVM (libSVM ou SMO). Em todos os testes foi utilizada *cross-validation* com 10 *folds* a partir dos bancos de dados descritos na Metodologia.

O primeiro algoritmo executado foi a libSVM. Todos os testes realizados com este algoritmo utilizaram o tipo C-SVC de SVM, que é utilizado como classificador com custo $C = 1$. O primeiro *kernel* utilizado foi o chamado *Radial Basis Function* (RBF), que funciona como uma medida de similaridade entre os dados, sendo baseado na distância Euclidiana [31] e cuja formulação pode ser vista na Equação 4. Cinco valores foram utilizados para o parâmetro γ, a fim de obter melhor adequação: 0 (interpretado pelo Weka como $\frac{1}{maxIndex}$), 1, 0,1, 0,01, 0,5. Para o primeiro experimento, este método obteve coeficiente Kappa 0,8417, com 89,4% de taxa de acerto com $\gamma = 0$, 0,8083 e 87,22% com $\gamma = 1$, 0,8833 e 92.22% com $\gamma = 0,1$, 0,7750 e 85% com $\gamma = 0,01$ e 0,8750 e 91,66% com $\gamma = 0,5$. Já para o segundo experimento, foi obtido um Kappa de 0,03 com 35,6% de taxa de acerto para ($\gamma = 0$), 0 e 33,33% para ($\gamma = 1$), 0,07 e 37,78% para ($\gamma = 0,1$), 0,08 e 38,89% para ($\gamma = 0,01$) e 0,1 e 40% para ($\gamma = 0,5$).

$$f_i = \exp^{-\gamma|u-v|^2} \tag{4}$$

Continuando com a libSVM, mas agora com o *kernel* linear (que significa, concre-

tamente, que não se está utilizando um *kernel*, mas resolvendo o problema de otimização original do método, mostrado na Equação 1), foram obtidos Kappas 0,3917 e 0,2, com taxas de acerto 59,4% e 46,7% para os experimentos 1 e 2, respectivamente.

O último *kernel* executado junto do método libSVM foi o polinomial, que assim como o RBF também é utilizado como medida de similaridade, sendo a Equação 5 sua representação na libSVM, com c_0 representando um coeficiente constante e d o grau do polinômio. O valor utilizado para c_0 foi 1, para d foi 3 e os valores de γ permanecem os mesmos descritos para o RBF. As taxas de acerto passaram a ser 81.1% ($\gamma = 0$), 60% ($\gamma = 1$), 70,55% ($\gamma = 0,1$), 86,66% ($\gamma = 0,01$) e 63,33% ($\gamma = 0,5$) para o primeiro experimento e 73,89%, 73,33%, 72,22%, 73,33% e 73,33, respectivamente, para o segundo, com coeficientes Kappa de 0,7167, 0,4, 0,5583, 0,8 e 0,45 para os testes com o primeiro experimento e 0,6083, 0,60, 0,5833, 0,60 e 0,60 respectivamente, para o experimento 2.

$$f_i = (\gamma * u^T * v + c_0)^d \tag{5}$$

Passando-se, então, para o método SMO, temos a execução de quatro outros *kernels* com a constante de custo sendo mantida em 1, assim como foi feito com a libSVM. O primeiro *kernel* executado foi o polinomial fornecido pelo Weka que, diferente do anterior, não permite variação de γ, apenas de grau, tendo esta opção recebido valor 3. O primeiro experimento gerou os valores 0,7083 e 80,55% como Kappa e taxa de acerto, respectivamente, e o segundo experimento gerou os valores 0,6333 e 75,56%.

O próximo *kernel* executado em conjunto com o SMO foi o *kernel* polinomial normalizado, que tem as mesmas características do *kernel* polinomial anterior, mas que normaliza os vetores de dados para a execução dos cálculos. Para o primeiro experimento obteve-se Kappa de 0,5417 e a taxa de acerto 69,44%; e para o segundo experimento forneceu coeficiente 0,5083 e taxa 67,22%.

Em seguida, foi utilizado o *kernel* Puk junto do método SMO. Este *kernel* utiliza a função Pearson VII, que tem formato similar à função normal, sendo representada pela Equação 6 [32]. Dois parâmetros podem ser variados neste kernel, σ e ω. Foram feitos os seguintes pares para teste: (1,1), (0,1, 1), (1, 0,1) e (0,1, 0,1). Ao executar o experimento 1, o Weka retornou em seu relatório o valor de coeficiente Kappa 0,8833 e taxa de acerto de 92,2% para o primeiro par de parâmetros, 0,8667 e 91,11% para o segundo, 0,8917 e 92,78% para o terceiro e 0,85 e 90% para o último par. Já ao executar o segundo experimento, o coeficiente Kappa obtido foi 0,8917, com taxa de acerto de 92,8% para o primeiro par, 0,8250 e 88,33% para o segundo, 0,7917 e 86,11% para o terceiro e 0,8333 e 88,89% para o quarto par.

$$f_i = \frac{1}{[1 + (2\sqrt{\|u - v\|^2}\frac{\sqrt{2^{(1/\omega)}-1}}{\sigma})^2]^\omega} \tag{6}$$

Finalmente, o último *kernel* utilizado foi o RBF, mas assim como o polinomial, com o SMO utilizou-se a versão fornecida pelo Weka, que também permite alterar os valores para γ e, por isso, foram utilizados os mesmos valores estabelecidos anteriormente para estes novos testes, com exceção do 0, já que este *kernel* não faz a mesma interpretação que o seu equivalente na libSVM faz. Foram observados coefiecientes Kappa 0,8417 ($\gamma = 1,0$), 0,60 ($\gamma = 0,1$), 0,175 ($\gamma = 0,1$) e 0,8250 ($\gamma = 0,5$) no primeiro experimento e 0,8750, 0,6167, 0,15 e 0,85, respectivamente, no segundo experimento. As taxas de acerto observadas no primeiro experimento foram 89,44%, 73,33%, 45% e 88,33% respectivamente. Já no segundo experimento, foram observadas, respectivamente, as seguintes taxas de acerto: 91,67%, 74,44%, 43,3% e 90%. Os melhores resultados obtidos com cada *kernel* para cada experimento são sintetizados na Tabela 2.

A Tabela 2 inclui ainda os resultados obtidos com os algoritmos J48 e JRip disponíveis também no *software* Weka, que representam os modelos árvore de decisão e lógica clássica, respectivamente. Com o algoritmo J48 foi obtida taxa de acerto de 96,11% e Kappa 0,9417 para o experimento 1 e taxa 87,78% e Kappa 0,8167 para o experimento 2. Já o algoritmo JRip apresentou taxas de acerto 86,11% e 88,33% e coeficientes Kappa 0,7917 e 0,8250 para os experimentos 1 e 2, respectivamente.

Dentre todos os diferentes *kernels* utilizados, o Puk foi o que retornou melhor coeficiente Kappa para ambos os experimentos. É interessante observar também o mau desempenho do *kernel* RBF com a libSVM no segundo experimento, o que indica que a implementação do *kernel* RBF utilizada pelo SMO é mais estável, em relação a estes conjuntos de dados. Nota-se, no entanto, que este *kernel* parece ser adequado para o experimento 1, principalmente no resultado obtido com a libSVM. Isto é esperado, uma vez que é possível separar os pontos iniciais e finais em regiões, e as distâncias entre tais regiões são bem-definidas. Outro *kernel* que obteve bons resultados em ambos os experimentos foi o polinomial, mas ele não se adequa a sistemas de avaliação em tempo real, pois ele demorou vários minutos para terminar seus cálculos. O *kernel* linear não obteve resultados satisfatórios em nenhum dos dois experimentos, o que era esperado, uma vez que os dados utilizados, especialmente no experimento 2, são não-lineares. É notável também que quando comparada a outros métodos, especialmente no experimento 2, a SVM com *kernel* Puk ainda apresenta melhores resultados, confirmando a ideia de que a SVM é adequada para a tarefa de avaliação de corte, em especial de trajetória.

8 Considerações Finais

A partir do exposto neste artigo, é possível perceber que tanto o processo de corte virtual quanto o de avaliação de usuários de simuladores são complexos, sendo a integração de ambos em um único sistema uma tarefa árdua. É necessário, então, fazer o planejamento e

Tabela 2. Coeficiente Kappa e taxa de acerto obtidos com os métodos libSVM e SMO e 7 diferentes kernels

	Coeficiente Kappa	Taxa de Acerto
libSVM com kernel RBF		
Experimento 1	0,8833	92,22
Experimento 2	0,1000	40,00
libSVM com kernel linear		
Experimento 1	0,3917	59,40
Experimento 2	0,2000	46,70
libSVM com kernel polinomial		
Experimento 1	0,8000	86,66
Experimento 2	0,6083	73,89
SMO com kernel polinomial		
Experimento 1	0,7083	80,55
Experimento 2	0,6333	75,56
SMO com kernel polinomial normalizado		
Experimento 1	0,5417	69,44
Experimento 2	0,5083	67,22
SMO com kernel Puk		
Experimento 1	0,8917	92,78
Experimento 2	0,8917	92,80
SMO com kernel RBF		
Experimento 1	0,8417	89,44
Experimento 2	0,8750	91,67
J48		
Experimento 1	0,9417	96,11
Experimento 2	0,8167	87,78
JRip		
Experimento 1	0,7917	86,11
Experimento 2	0,8250	88,33

o levantamento de requisitos de ambos com atenção, para que o processo de implementação possa ser conduzido de maneira coordenada.

A escolha da técnica de corte que será utilizada (incluindo a escolha da representação geométrica, da técnica de detecção de colisão e da técnica de deformação) depende bastante de procedimento cirúrgico que será simulado e das características desejadas no sistema, como realismo físico, alto desempenho, alta qualidade gráfica, etc. Já a escolha do modelo de decisão que será utilizado na avaliação dos usuários deve ser feita com base nos parâmetros que serão utilizados para avaliar. No caso específico das incisões, os sistemas baseados em lógica *fuzzy* podem se mostrar adequados para esta tarefa, uma vez que eles são capazes de descrever os tipos de variáveis presentes no problema e de acomodar as imprecisões. Apesar disso, outros modelos que não utilizam lógica *fuzzy* também podem ser capazes de se adequar ao problema da avaliação do corte, como é o caso das SVMs analisadas aqui. Apesar de dois *kernels* terem retornado resultados insatisfatórios (com coeficientes Kappa baixos), o método libSVM com *kernel* polinomial e com o *kernel* RBF (para o experimento 1) e o método SMO com os *kernels* Puk, polinomial e RBF se mostram bons candidatos para uso em tarefas de classificação relacionadas ao corte, ao menos no que diz respeito à trajetória do corte, sendo esta análise a principal contribuição deste trabalho.

É importante notar que a maioria dos simuladores apresentados neste artigo não apresentam suporte à avaliação de usuário, sendo muitos nem mesmo validados por especialistas. Estes são aspectos importantes nesse contexto, especialmente pela intenção de se utilizar os simuladores para treinar futuros profissionais. É compreensível, no entanto, que haja essa lacuna nos trabalhos da área, uma vez que a inexistência de métricas universalmente aceitas pela comunidade científica representa um grande empecilho para a realização destas tarefas, o que representa, inclusive, uma limitação para este trabalho.

Finalmente, destacam-se como trabalhos futuros a oportunidade de testar o poder de avaliação das SVMs com outras variáveis importantes do corte, como a profundidade, a força aplicada, etc., a comparação das classificações obtidas com a SVM com outras obtidas utilizando outros modelos, o uso de dados obtidos a partir de outros dispositivos de interação para geração dos conjuntos de teste e a validação destes dados por especialistas, que poderiam transformar os dados em algo que se assemelha mais à realidade.

Contribuição dos autores:

- Ives Fernando Martins Santos de Moura: Responsável pela execução dos experimentos descritos no artigo. Participou ainda da escrita de todas as seções do texto.

- Ronei Marcos de Moraes: As seções de avaliação, do método SVM e elaboração dos experimentos foram construídas com contribuições deste autor. Também foi realizada a

revisão final do texto.

- Liliane dos Santos Machado: A introdução do artigo e as seções associadas aos conceitos de corte virtual e simuladores com Realidade Virtual foram elaboradas com contribuições desta autora.

Referências

[1] Moraes RM, Machado LS. Assessment Systems for Training Based on Virtual Reality: A Comparison Study. SBC Journal on 3D Interactive Systems. 2012;3(1):9–16.

[2] Moraes RM, Rocha AV, Machado LS. Intelligent assessment based on Beta Regression for realistic training on medical simulators. Knowledge-Based Systems. 2012;32:3–8.

[3] Wu J, Westermann R, Dick C. Physicallybased simulation of cuts in deformable bodies: A survey. Eurograph State-of-the-Art Report. 2014;2.

[4] Pan JJ, Chang J, Yang X, Liang H, Zhang JJ, Qureshi T, et al. Virtual reality training and assessment in laparoscopic rectum surgery. The International Journal of Medical Robotics and Computer Assisted Surgery. 2014;.

[5] SenseGraphics. References; 2016. http://sensegraphics.com/references/.

[6] Patnaik V, Singla RK, Bansal V. Surgical incisions - their anatomical basis Part IV - abdomen. J Anat Soc India. 2001;50(2):170–8.

[7] Tang M, Manocha D, Lin J, Tong R. Collision-streams: fast gpu-based collision detection for deformable models. In: Symposium on interactive 3D graphics and games. ACM; 2011. p. 63–70.

[8] Bruyns CD, Senger S, Menon A, Montgomery K, Wildermuth S, Boyle R. A survey of interactive mesh-cutting techniques and a new method for implementing generalized interactive mesh cutting using virtual tools. The journal of visualization and computer animation. 2002;13(1):21–42.

[9] Choi KS. Interactive cutting of deformable objects using force propagation approach and digital design analogy. Computers & Graphics. 2006;30(2):233–243.

[10] Wu J, Westermann R, Dick C. Real-time haptic cutting of high resolution soft tissues. Studies Health Technol Inform(Proc Medicine Meets Virtual Reality 2014). 2014;196:469–475.

[11] Jin X, Joldes GR, Miller K, Yang KH, Wittek A. Meshless algorithm for soft tissue cutting in surgical simulation. Computer methods in biomechanics and biomedical engineering. 2014;17(7):800–811.

[12] Ericson C. Real-time collision detection. CRC Press; 2004.

[13] Zhang Y, Yuan Z, Ding Y, Zhao J, Duan Z, Sun M. Real Time Simulation of Tissue Cutting Based on GPU and CUDA for Surgical Training. In: Biomedical Engineering and Computer Science (ICBECS), 2010 International Conference on. IEEE; 2010. p. 1–4.

[14] Lu Z, Arikatla VS, Han Z, Allen BF, De S. A physics-based algorithm for real-time simulation of electrosurgery procedures in minimally invasive surgery. The International Journal of Medical Robotics and Computer Assisted Surgery. 2014;10(4):495–504.

[15] Pan J, Bai J, Zhao X, Hao A, Qin H. Real-time haptic manipulation and cutting of hybrid soft tissue models by extended position-based dynamics. Computer Animation and Virtual Worlds. 2015;26(3-4):321–335.

[16] Sela G, Subag J, Lindblad A, Albocher D, Schein S, Elber G. Real-time haptic incision simulation using FEM-based discontinuous free-form deformation. Computer-Aided Design. 2007;39(8):685–693.

[17] Ho AK, Alsaffar H, Doyle PC, Ladak HM, Agrawal SK. Virtual reality myringotomy simulation with real-time deformation: Development and validity testing. The Laryngoscope. 2012;122(8):1844–1851.

[18] Xia P, Lopes AM, Restivo MT. Virtual reality and haptics for dental surgery: a personal review. The Visual Computer. 2013;29(5):433–447.

[19] Moraes RM, Machado LS. Psychomotor skills assessment in medical training based on virtual reality using a Weighted Possibilistic approach. Knowledge-Based Systems. 2014;70:97–102.

[20] Paiva PVF, Machado LS, Gondim Valença MM. A Virtual Environment for Training and Assessment of Surgical Teams. In: XV Symposium on Virtual and Augmented Reality (SVR). IEEE; 2013. p. 17–26.

[21] Wiet G, Hittle B, Kerwin T, Stredney D. Translating surgical metrics into automated assessments. Studies in health technology and informatics. 2012;173:543.

[22] Paiva PVF, Machado LS, Valença AMG, Moraes RM. Collaborative Evaluation of Surgical Team's Training Based on Virtual Reality; 2015. Manuscript.

[23] Santos AD, Machado LS, Moraes RM, Gomes RG. Avaliação baseada em lógica fuzzy para um framework voltado à construção de simuladores baseados em RV. In: Anais do XII Symposium on Virtual and Augmented Reality. Natal: Sociedade Brasileira de Computação; 2010. p. 194–202.

[24] Allen B, Nistor V, Dutson E, Carman G, Lewis C, Faloutsos P. Support vector machines improve the accuracy of evaluation for the performance of laparoscopic training tasks. Surgical endoscopy. 2010;24(1):170–178.

[25] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273–297.

[26] Fan RE, Chen PH, Lin CJ. Working set selection using second order information for training support vector machines. The Journal of Machine Learning Research. 2005;6:1889–1918.

[27] NG A. Support Vector Machine - Optimization Objective;. Acessado em: 2015-12-16. https://class.coursera.org/ml-003/lecture.

[28] Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. The annals of statistics. 2008;p. 1171–1220.

[29] NG A. Support Vector Machine - Kernels I;. Acessado em: 2015-12-16. https://class.coursera.org/ml-003/lecture.

[30] Platt J, et al. Fast training of support vector machines using sequential minimal optimization. Advances in kernel methods - support vector learning. 1999;3.

[31] Vert JP, Tsuda K, Schölkopf B. A primer on kernel methods. Kernel Methods in Computational Biology. 2004;p. 35–70.

[32] Abakar KA, Yua C. Performance of SVM based on PUK kernel in comparison to SVM based on RBF kernel in prediction of yarn tenacity. Indian Journal of Fibre & Textile Research. 2014;39:55–59.

Estudo da Relação Entre o Gênero de Espectadores e o Conteúdo de Vídeos

Study of the Relationship Between Spectators Gender and Video Content

Ítalo de Pontes Oliveira [1, 2, 3]
Eanes Torres Pereira [1, 4]
Herman Martins Gomes [1, 5]

Data de submissão: 15/09/2015, *Data de aceite: 10/05/2016*

Resumo: A sinalização digital pode ser vista como uma forma de publicidade exibida painéis posicionados em locais com circulação ou presença de pessoas, a qual pode ser incrementada com métodos de visão computacional para oferecer recursos de medição da audiência. Neste contexto, o presente trabalho apresenta a avaliação experimental de uma abordagem para estimar, a partir de vídeos, o perfil de espectadores relacionando o gênero e os conteúdos exibidos. A abordagem foi implementada utilizando o detector de faces proposto por Viola-Jones e um classificador de gênero com base em características *Fisherfaces*, validados utilizado a base de faces FERET. O foco central deste trabalho foi o levantamento e teste experimental de hipóteses relativas à associação entre o gênero e categorias de conteúdos exibidos em sinalização digital. A análise realizada a partir de dados obtidos de um ambiente de cantina, em que os espectadores, monitorados por uma câmera de vídeo, assistiam a conteúdos selecionados em uma TV enquanto se alimentavam, mostrou que existe diferença estatística na atenção dada por homens e mulheres em função do tipo de conteúdo. Mais especificamente, (i) homens assistiram mais aos anúncios do que as mulheres; e (ii), os homens e as mulheres se concentraram mais aos anúncios do tipo esportivo do que aos anúncios do tipo jornalístico.

***P*alavras-chave:** sinalização digital, detecção de faces, classificação de gênero

[1]Laboratório de Percepção Computacional (LPC), Departamento de Sistemas e Computação (DSC), Centro de Engenharia Elétrica e Informática (CEEI), Universidade Federal de Campina Grande (UFCG). R. Aprígio Veloso, 882 - Bairro universitário, Campina Grande - PB, 58429-900.
[2]Bolsista do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)
[3]{italooliveira@copin.ufcg.edu.br}
[4]{eanes@computacao.ufcg.edu.br}
[5]{hmg@dsc.ufcg.edu.br}

Abstract: Digital signage can be viewed as a form of advertising exibited in displays positioned in spaces with transit or presence of people which can be enhanced with computer vision methods to offer resources of measuring the audience. In this context, the present work shows the experimental evaluation of an approach for estimating, from videos, the profile of viewers by relating gender and the displayed content. The approach was implemented using a face detector proposed by Viola-Jones and a genre classifier based on Fisherfaces features, which were validated using the FERET face dataset. The central focus of this study was to formulate and to experimentally test hypotheses regarding the association between gender and content categories displayed in digital signage. The analysis of data obtained from a cafeteria environment where viewers, being monitored by a video camera, watched selected contents on a TV whilefeeding, showed that there is an statistical differece in the attention of men and women with regards to the type of displayed contents. More specifically, (i) men payed more attention to news programs than women; and (ii), men and women focused more on sports contents than on news contents.

***K*eywords:** digital signage, face detection, gender classification

1 Introdução

A sinalização digital expõe informações como notícias urgentes, previsão do tempo, entretenimento, etc. para captação da atenção das pessoas, empregando telas de LCD ou plasma de alta resolução, com seus conteúdos atualizados em tempo real [15]. Tem sido utilizada para diversas finalidades, tais como: (i) *marketing*, na promoção e lançamento de produtos e serviços; (ii) anúncios, vendendo tempo de transmissão para fornecedores ou terceiros; (iii) levar informação aos frequentadores de um determinado ambiente; (iv) *infotainment*, cujo propósito é o entretenimento para reduzir a percepção do tempo de espera; dentre outros. Os contextos de aplicação são diversos, incluindo bancos, aeroportos, academias, igrejas, hospitais/consultórios médicos, feiras, etc.

Grandes empresas utilizam e investem em sinalização digital para a divulgação de marcas ou mercadorias, ao mesmo tempo que desenvolvem novas tecnologias que promovem melhor experiência do usuário, provendo maior interatividade, a exemplo de Google® [20] e Intel® [4]. Padrões também tem sido definidos, como o *Point of Purchase Advertising International* (POPAI) [15], apoiado por empresas como Cisco® e LG®.

O problema de negócio ao qual este trabalho está associado consiste em conhecer o perfil dos espectadores e usar essa informação para otimizar a audiência, visando maximizar o *Return On Investment* (ROI) dos anunciantes. Câmeras podem ser utilizadas para realizar o e monitoramento dos espectadores. A Intel®, por exemplo, desenvolveu um Arcabouço

de Publicidade Inteligente (API) integrando Análise de Observadores Anônimos (AOA) e mineração de dados, para transmitir em tempo real o anúncio mais adequado ao público presente, provendo interatividade de acordo com o perfil dos espectadores [18].

Want e Schilit [20], consideram que a interatividade poderá ser uma peça chave na sinalização digital visto que, estima-se que o mercado receberá US$ 5 bilhões em investimentos até 2016 e até então, nenhum modelo de interatividade está bem concebido. Um desafio relevante desta área é o desenvolvimento de métodos para traçar o perfil dos espectadores a partir da análise de vídeo por meio de técnicas como a detecção de faces, a classificação de gênero e idade, dentre outros. O propósito é lançar anúncios personalizados para atingir o público-alvo de maneira mais efetiva e ajustar os conteúdos exibidos em tempo real.

Tratar os consumidores de forma personalizada é motivo de sucesso para algumas empresas, como é o caso das empresas de e-*commerce* que utilizam os *cookies* dos navegadores para personalizar o atendimento dos clientes. A Marisa®, contratou a empresa AlmapBBDO® para desenvolver um aplicativo com anúncios interativos voltado para o público masculino, com versões para Android e iPad [2]. Em um mês, o *e-commerce* da Marisa® obteve oito mil novos consumidores e a AlmapBBDO® recebeu o prêmio leão de bronze no Festival de Cannes 2014 na categoria de publicidade multimídia. A partir dos fatos relatados, pode-se observar que combinar interatividade e conhecimento do perfil dos clientes/consumidores/espectadores tem forte potencial em aplicações comerciais.

Inserido no contexto apresentado, o presente trabalho apresenta um sistema para análise de atenção dos espectadores, e um estudo de caso em um cenário real. A atenção é analisada indiretamente a partir de evidência proveniente de um detector de faces. Neste trabalho, o perfil dos espectadores é mensurado com respeito à variável gênero, estimada a partir de um classificador treinado com imagens de faces. A principais contribuições do trabalho são as seguintes:

- Avaliação experimental de um sistema para análise da atenção dos espectadores fundamentado em código aberto (*Open Source Computer Vision Library* (OpenCV)[6]), formado por um detector de faces (Viola-Jones) [19] e um classificador de gênero (Fisherfaces) [14]. Foram avaliados o impacto do tamanho do conjunto de treinamento no tempo de processamento e na acurácia do classificador de gênero.

- Estudo de caso envolvendo a análise da atenção de espectadores, a qual foi realizada transmitindo anúncios esportivos e jornalísticos em uma cantina, em que foram testadas hipóteses sobre a preferência a categorias de vídeos em relação ao gênero dos espectadores.

[6]Disponível em `http://opencv.org`, acessado em abril de 2016.

As conclusões deste trabalho, a partir de uma análise estatística sobre as estimativas de gênero obtidas durante a exibição dos vídeos, indicam que existe diferença estatística na atenção dada por homens e mulheres em função do tipo de conteúdo. A análise mostrou que: (i) homens assistiram mais aos anúncios do que as mulheres; e (ii), os homens e as mulheres se concentraram mais aos anúncios do tipo esportivo do que aos anúncios do tipo jornalístico.

O artigo está organizado como descrito a seguir. A Seção 2 apresenta uma discussão sobre os trabalhos relacionados. As questões de pesquisa e a definição das hipóteses são feitas na Seção 3 e os materiais e métodos utilizados são descritos na Seção 4. Na Seção 5, é apresentado um estudo sobre o tempo de processamento para o treinamento do classificador de gênero. Uma análise sobre a acurácia do classificador se encontra na Seção 6. A análise dos vídeos capturados na catina é apresentada na Seção 7. Finalmente, as conclusões do trabalho e propostas para trabalhos futuros são apresentadas na Seção 8.

2 Trabalhos Relacionados

No trabalho de Ravnik [9] foi utilizada uma câmera monocular para segmentação dos espectadores. O algoritmo para detecção de faces foi implementado utilizando a biblioteca de código aberto OpenCV, facilitando a replicação do experimento. Além disso, foram comparados dois tipos de câmera: uma *webcam* logitech Pro 9000 e uma câmera Kinect RGB-D. Foi observado que para distâncias acima de quatro metros, o classificador de gênero obteve melhores resultados com os vídeos da *webcam*. No estudo de caso realizado no presente trabalho foi utilizada uma *webcam* do mesmo modelo, uma vez que os códigos implementados no OpenCV obtiveram melhores resultados nesta câmera.

O sistema proposto em [9] foi aplicado a cenário real de sinalização digital [10], localizado em uma boutique de roupas em Ljubljana, capital da Eslovênia. A atenção dada pelos espectadores foi correlacionada ao tipo de conteúdo transmitido, ao grupo de idade e ao tempo de atenção. Entretanto, não foi apresentada qualquer análise estatística quanto à normalidade dos dados e outros parâmetros que eventualmente causariam ameaças à validade. Similarmente à pesquisa reportada em [9], no presente trabalho também foram realizadas análises da atenção dos espectadores, entretanto, foi realizada uma validação mais robusta do sistema que gera os dados experimentais, fato que aumentou a consistência das inferências. Em [11], o tempo de atenção dos espectadores foi correlacionado ao gênero, sendo verificado que o tempo médio de atenção dos homens (1,2s) foi substancialmente maior do que o tempo médio das mulheres (0,4s), resultado alinhado ao que foi observado no presente trabalho (ver Seção 7).

No trabalho de Yin *et. al.* [5], foram estimados o tempo médio de visualização e o número de pessoas que assistiram a diferentes tipos de anúncios, com o propósito de definir o perfil dos espectadores. Naquele trabalho, observou-se uma relação inversa quanto ao tempo

de atenção e o número de pessoas que assistiram às notícias globais e locais, i.e., enquanto as notícias locais atraíram um número maior de espectadores do que notícias globais, as notícias globais receberam tempo de atenção do telespectador inversamente proporcional à quantidade de expectadores presentes. Para outros tipos de anúncios, os tempos de atenção não apresentaram tais variações. Entretanto, estas análises se limitaram ao tipo de conteúdo transmitido e ao tempo de visualização, não sendo apresentadas considerações demográficas, tais como relacionar o conteúdo ao gênero (que foi realizado no presente trabalho) ou à idade do espectador.

O presente trabalho preenche lacunas nas pesquisas supracitadas ao apresentar uma validação experimental do classificador de gênero, que fornecerá os dados necessários para análise da atenção dos espectadores. Esse tipo de validação não é observada em outros trabalhos relacionados [9], [10], [11]. Além disso, a abordagem foi implementada utilizando código-fonte aberto, permitindo replicação das análises feitas. Por fim, correlacionar o gênero do espectador ao conteúdo tem se mostrado relevante para que as empresas atinjam seu público-alvo com maior eficácia [5].

3 Questões de Pesquisa e Definição de Hipóteses

Antes de realizar análises da atenção dos espectadores, realizou-se uma validação do classificador de gênero, o qual gerou os dados de entrada para as inferências. Neste sentido, foram formuladas questões de pesquisa associadas à relação entre a quantidade de imagens no conjunto de treinamento e o tempo de processamento, e à acurácia [16] do classificador de gênero.

Foram realizados testes paramétricos para determinar se o tempo de processamento para os diferentes conjuntos de treinamento obedecem a uma distribuição normal e para determinar se os tempos de processamento para os diferentes conjuntos de treinamento são iguais. A hipótese formulada é que, quanto maior o conjunto de treinamento, maior será o tempo de processamento. Por fim, foi testado se, ao aumentar o número de imagens no conjunto de treinamento, ocorre um aumento na acurácia do classificador de gênero. Desta forma, foi realizado o cálculo dos índices de correlação entre as medidas de acordo com o tipo de distribuição previamente concebido.

Mais especificamente, as seguintes questões de pesquisa foram formuladas:

$\mathbf{Q}_1$: Os tempos para treinar classificadores de gênero, com conjuntos de treinamento de diferentes tamanhos, são iguais?

$\mathbf{Q}_2$: Os tempos de treinamento obedecem a uma distribuição normal?

$\mathbf{Q}_3$: O tamanho do conjunto de treinamento afeta a acurácia do classificador de gê-

nero?

Na Tabela 1, podem ser observados os testes de hipóteses realizados neste trabalho com respeito ao classificador de gênero. O superconjunto **C** contém conjuntos de treinamento. Ou seja, c_i é o conjunto de treinamento i, $N(\mu, \sigma^2)$ indica uma distribuição normal, com média μ e variância $\sigma^2 > 0$. A função $t(c_i)$ retorna o tempo de processamento para o conjunto de treinamento c_i. Por fim, a função $corr(\ )$ retorna a correlação com a acurácia definida por ACC.

Tabela 1. Matriz de hipóteses considerando o tempo de processamento e o conjunto de treinamento.

Hipóteses Nulas	**Hipóteses Alternativas**
$\mathrm{H}_{1\_1} : \forall c_{i,j} \exists C_{i \neq j}\ t(c_i) = t(c_j)$	$\mathrm{H}_{1\_2} : \forall c_{i,j} \exists C_{i \neq j}\ t(c_i) \neq t(c_j)$
$\mathrm{H}_{2\_1} : \forall c \exists C\ f(t(c)) = N(\mu, \sigma^2)$	$\mathrm{H}_{2\_2} : \forall c \exists C\ f(t(c)) \neq N(\mu, \sigma^2)$
$\mathrm{H}_{3\_1} : \lvert corr(\text{ACC, Tam})\rvert \geq 0,7$	$\mathrm{H}_{3\_2} : \lvert corr(\text{ACC, Tam})\rvert < 0,7$

A questão de pesquisa central desta investigação, no contexto da análise da atenção dos expectadores, é a seguinte:

$\mathbf{Q}_4$: Homens e mulheres prestam atenção igualmente a anúncios jornalísticos e esportivos?

As variáveis envolvidas no experimento projetado para responder à pergunta acima são apresentadas na Tabela 2. O experimento considera dois fatores com dois níveis cada. Isto é, os gêneros possíveis são masculino e feminino, enquanto que o tipo de anúncio transmitido pode ser esportivo ou jornalístico. Sendo assim, a base das análises consistirá na combinação deste fatores.

Tabela 2. Variáveis analisadas.

Fator	**Níveis**
Gênero	Masculino
	Feminino
Tipo de anúncio	Esportivo
	Jornalístico

Por fim, na Tabela 3, pode ser observada uma matriz de hipóteses em que na coluna da esquerda são apresentadas as hipóteses nulas e, na coluna da direita as hipóteses alternativas associadas à questão de pesquisa $\mathbf{Q}_4$. A variável μ representa a média da quantidade de espectadores, os índices M e F rotulam o gênero em Masculino ou Feminino, respectivamente.

Enquanto os sobrescritos e e j, indicam o tipo do conteúdo transmitido em vídeos esportivos e jornalísticos, respectivamente. Por exemplo, μ^e_M é a média de homens detectados que assistiram ao anúncio esportivo.

Tabela 3. Matriz de hipóteses considerando o gênero do telespectador e o tipo do conteúdo transmitido.

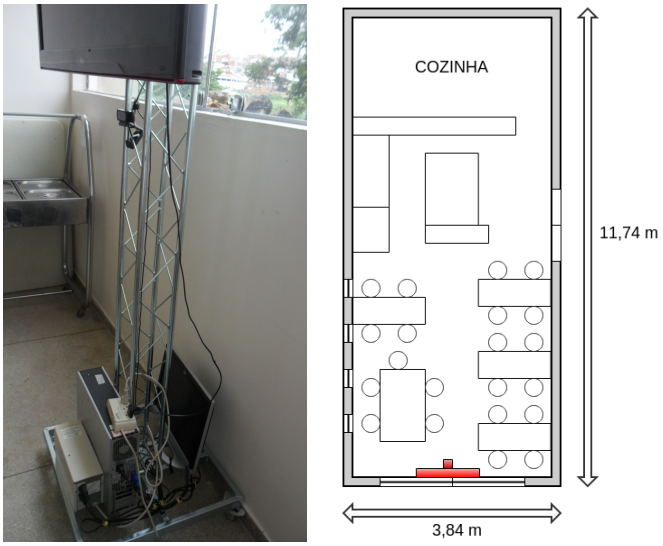
Hipóteses Nulas	**Hipóteses Alternativas**
H_{3_1}: $\mu^j_M = \mu^e_M$	H_{3_2}: $\mu^j_M \neq \mu^e_M$
H_{4_1}: $\mu^j_F = \mu^e_F$	H_{4_2}: $\mu^j_F \neq \mu^e_F$

4 Materiais e Métodos

O estudo de campo do sistema de sinalização digital, para análise da atenção dos espectadores correlacionado o gênero ao tipo de conteúdo, foi realizado em uma cantina localizada no campus sede da Universidade Federal de Campina Grande, Paraíba, a qual possui mais de sete mil discentes e mais de 800 docentes vinculados. Para sinalização digital, foi utilizado um televisor LG 42" com resolução *Full* HD (1920×1080) e formato de tela 16:9. Tal equipamento está de acordo com o padrão definido no POPAI [8] para a exibição de vídeos na sinalização digital. A câmera utilizada para captura dos espectadores foi a Logitech HD Pro *Webcam* C920 com gravação *Full* HD codec H.264 e 24 fps. Os algoritmos foram desenvolvidos na linguagem de programação C/C++, usando a biblioteca OpenCV 2.4.8 com *GNU Compiler Collection* (GCC) versão 4.8.2 executados em uma máquina Intel® Core™ i7 3770 3,40GHz com 8GB DDR3 de memória principal e placa de vídeo nVIDIA 210/1GB com GNU/Linux Ubuntu 14.04 LTS 64 *bits*.

Na Figura 1, é apresentado o ambiente do experimento. Na Subfigura 1(a), pode ser observada a estrutura montada para transmissão da sinalização digital e a captura do vídeo dos espectadores e a Subfigura 1(b) representa a planta baixa do estabelecimento comercial ($\sim$45m^2), em que o equipamento de transmissão da sinalização é destacado em vermelho, situado frontalmente aos espectadores que utilizam o local.

Na Figura 2, pode ser observado o diagrama de fluxo que representa o funcionamento lógico da aplicação desenvolvida para realizar a análise da atenção dos espectadores. Os vídeos capturados das pessoas no estabelecimento comercial são processados quadro-a-quadro, as faces dos vídeos são detectadas automaticamente pelo OpenCV usando o algoritmo de Viola e Jones [19] e então equalizadas, pois segundo Khan *et. al.* [12] o classificador de gênero apresenta melhores resultados após a equalização dos histogramas. Em seguida, o classificador de gênero previamente treinado, recebe a face por parâmetro para determinação do gênero. Após isso, o contador de gênero é atualizado, i.e., as quantidades das faces detectadas



(a) (b)

Figura 1. Estrutura e cenário da realização do experimento: (a) equipamento utilizado na sinalização digital; (b) planta baixa do local.

para cada gênero no quadro são armazenadas para posterior análise.

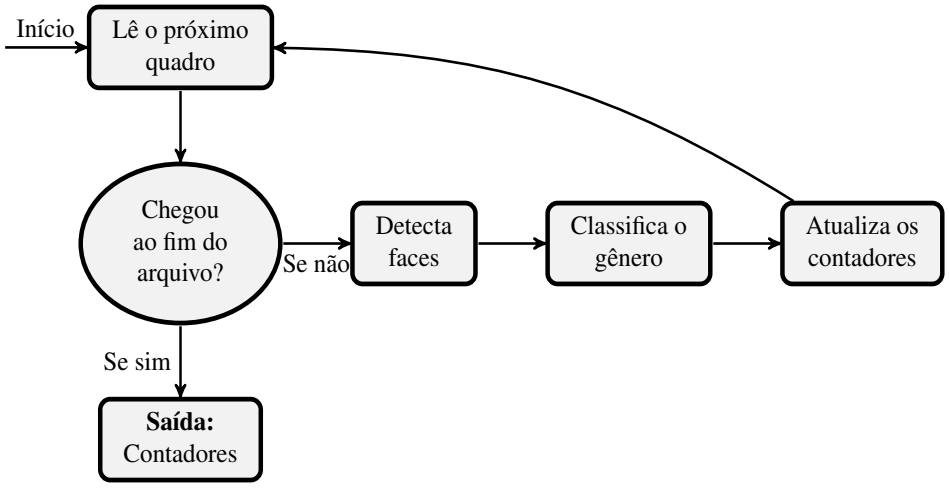


Figura 2. Diagrama de fluxo da aplicação.

4.1 Considerações Éticas

O experimento está de acordo com a Resolução n° 466 do Conselho Nacional de Saúde [7], que define diretrizes e normas regulamentadoras de pesquisas envolvendo seres humanos. Os procedimentos adotados nesta pesquisa asseguraram a confidencialidade, a privacidade e a proteção da imagem dos espectadores.

5 Tempo de Treinamento do Classificador de Gênero

Para validar o classificador de gênero foi utilizada a base de dados FERET, em que 200 faces masculinas e 200 faces femininas foram selecionadas aleatoriamente para compor o grupo de imagens de teste. Foi utilizado o classificador de gênero Fisherfaces [14], disponível na biblioteca OpenCV. Alterou-se o tamanho do conjunto de treinamento para observar a variação da acurácia e do tempo de processamento para os seguintes cenários: C = {200, 250, 300, 350, 400, 450, 500, 550}, em que cada valor representa a quantidade de imagens no conjunto de treinamento. Para se determinar o intervalo de tempo, foi utilizada a estrutura de dados `timespec` do cabeçalho time.h da linguagem GNU C com precisão de nanosegundos, como pode ser observado no seguinte trecho de código:

```
1. struct timespec inicio={0,0}, fim={0,0};
2. clock_gettime( CLOCK_MONOTONIC, &inicio );
3. // <Colocar aqui> o trecho de código a ter
 // o tempo de processamento medido
4. clock_gettime( CLOCK_MONOTONIC, &fim );
5. double tempo_de_execucao;
6. tempo_de_execucao <<
 ((((double) fim.tv_sec + 1.0e-9*fim.tv_nsec)
 - (double) inicio.tv_sec + 1.0e-9*inicio.tv_nsec));
```

em que a variável `tempo_de_execucao` armazena o tempo decorrido entre o início e o final da execução do algoritmo.

Na Figura 3, pode ser observado o tempo médio de processamento considerando diferentes tamanhos do conjunto de treinamento, com intervalos de confiança de 95% calculados com a função `t.test( )$conf.int` da linguagem R. Assim, pode-se afirmar que houve diferença estatística para treinar com diferentes quantidades de imagens no grupo de treinamento, i.e., quanto mais imagens na base de treinamento, maior o tempo de processamento.

[7] Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/cns/2013/res0466_12_12_2012.html, acessado em abril de 2016.

Como consequência, pode-se rejeitar a hipótese nula H_{3_1}: $\mu_M^j = \mu_M^e$, ou seja, os tempos de treinamento para diferentes quantidades de imagens foram diferentes.

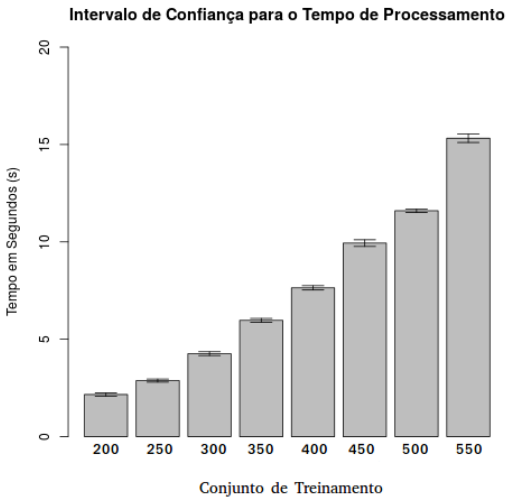


Figura 3. Intervalo de confiança de 95% para o tempo de treinamento do classificador de gênero, em função dos diferentes tamanhos de conjunto de treinamento.

Na Tabela 4, podem ser observados os resultados dos testes de normalidade Anderson-Darling [1] e Shapiro-Wilk [13], os quais evidenciam que os dados são normais, i.e., os valores obtidos foram menores do que 0,05. Por esta razão não é possível rejeitar a hipótese nula H_{0_1}: $\forall t(\ c_i\ ) := N(\mu, \sigma^2)$. A normalidade indica se os dados obedecem a uma distribuição normal, fato que ajuda a determinar quais os métodos estatísticos mais adequados ao problema.

Tabela 4. Teste de normalidade para os diferentes tamanhos de conjunto de treinamento do classificador de gênero.

Treinamento	**Anderson-Darling**	**Shapiro-Wilk**
200	1,2e-05	5,4e-06
250	2,2e-05	6,5e-06
300	7,8e-03	1,5e-05
350	3,9e-15	4,7e-13
400	2,2e-16	6,4e-16
450	6,2e-15	1,2e-10
500	1,8e-02	8,6e-03
550	2,2e-16	2,2e-16

Em seguida, foi utilizada a função `levene.test( )` do pacote "`car`" da linguagem R para teste de homocedasticidade, conforme proposto por John Fox [6], com intuito de verificar se os dados são estatisticamente iguais. Na Tabela 5, os resultados indicam que os dados são diferentes entre si, pois o Pr foi menor que o F-$Value$ com sete graus de liberdade, rejeitando-se, desta forma, a hipótese nula H_{0_1}: $\forall t(\ c_i\ ) := N(\mu, \sigma^2)$. Ou seja, o tempo de processamento foi diferente para as diferentes quantidades de imagens no conjunto de treinamento.

Tabela 5. Teste de homocedasticidade para o tempo de processamento do classificador de gênero.

Graus de Liberdade	**F-*Value***	**Pr(> F)**
7	19,9	2,2e-16

6 Análise Sobre a Acurácia do Classificador de Gênero

Neste experimento, como no anterior, foram alterados os tamanhos dos conjuntos de treinamento do classificador, mas agora o propósito foi avaliar o impacto dos diferentes tamanhos na acurácia (ACC) do classificador. A Equação 1 [16] define formalmente a acurácia.

$$\text{ACC} = \frac{VP + VN}{P + N}, \tag{1}$$

em que, VP corresponde à taxa de verdadeiros positivos. Convencionamos que o gênero masculino seria a classe positiva e o feminino, a negativa. Desta forma, VP corresponde à quantidade de homens classificados corretamente, VN corresponde à taxa de verdadeiros negativos, i.e., corresponde à quantidade de mulheres classificadas corretamente, P é a quantidade total de homens no conjunto de teste e N a quantidade total de mulheres do conjunto de teste.

A Figura 4 exibe a acurácia para os diferentes conjuntos de treinamento. Pode-se observar que há pouca variação estatística entre eles. Para verificar a variância da acurácia, foi realizado o `levene.test( )`. Como pode ser observado na Tabela 6, o valor calculado para a homocedasticidade foi menor que o F-$Value$. Portanto, pode-se rejeitar a hipótese nula H_{1_1}: $\forall t(\ c_i\ ) = t(\ \mu(\ \mathbf{C}\ )\ )$. Ou seja, existe diferença estatística para o valor da acurácia obtida utilizando-se diferentes tamanhos de conjunto de treinamento, assim como existe para o tempo de processamento (discutido na seção anterior). Por essa razão, faz-se necessário calcular a correlação entre a quantidade de imagens e a acurácia do classificador, com o objetivo de observar se ao se acrescentar imagens no conjunto de treinamento, ocorre aumento da acurácia ou vice-versa.

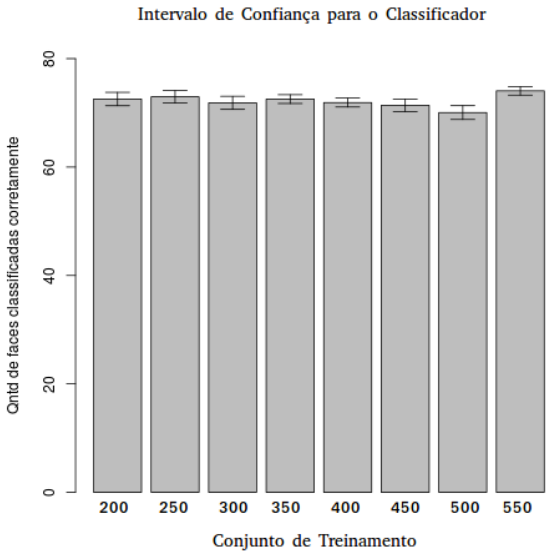


Figura 4. Intervalo de confiança de 95% para a acurácia do classificador de gênero.

Tabela 6. Teste de homocedasticidade para a acurácia.

Graus de Liberdade	**F-*Value***	**Pr(> F)**
7	4,8	2,318e-05

Para aferir se há relação entre a acurácia do classificador de gênero e a quantidade de imagens no conjunto de treinamento, podem ser utilizados os coeficientes de correlação. As seguintes medidas foram utilizadas como indicadores: Coeficiente de Correlação Linear de Pearson (CCLP) [7] e o Coeficiente de Correlação dos Postos de Spearman (CCPS) [17]. O CCLP permite identificar a linearidade entre dois conjuntos de medidas e assim, quanto maior o valor de CCLP, significa que há mais correlação entre o conjunto de treinamento e a acurácia. O CCLP entre dois conjuntos de medidas $A = \{a_k\}$ e $B = \{b_k\} \mid k = 1, 2, \ldots, K$ é dado por

$$\text{CCLP} = \frac{\sum_{k=1}^{K}(a_k - \bar{x}_a) \cdot (b_k - \bar{x}_b)}{\sqrt{\sum_{k=1}^{K}(a_k - \bar{x}_a)^2} \cdot \sqrt{\bar{x}_{k=1}^{K}(b_k - \bar{x}_b)^2}}, \tag{2}$$

em que $\bar{x}_a$ e $\bar{x}_b$ significam as médias amostrais das medidas dos conjuntos A e B, respectivamente.

O CCPS indica a dependência estatística entre dois conjuntos ordenados de medidas

Tabela 7. Correlação entre quantidade de imagens no conjunto de treinamento e a acurácia do classificador de gênero.

Qntd. de Imagens	CCLP	CCPS
200	-0,47	-0,55
250	-0,48	-0,57
300	-0,21	-0,31
350	-0,32	-0,47
400	-0,20	-0,41
450	-0,37	-0,49
500	0,10	-0,25
550	0,01	-0,30

A e B, sendo matematicamente expresso por

$$\text{CCPS} = 1 - \frac{6 \cdot \sum_{k=1}^{K} {d_k}^2}{K(K^2-1)} \tag{3}$$

em que K é o número de medidas e d_k é a diferença entre os postos das medidas ordenadas $a_k \in A$ e $b_k \in B$.

Na Figura 5, pode-se observar um gráfico de dispersão que relaciona a quantidade de imagens no conjunto de treinamento e a acurácia do classificador. Nota-se uma fraca tendência no canto superior direito de todos os gráficos, sugerindo que há pouca relação no aumento do número de imagens com a acurácia do classificador. Os resultados dos coeficientes de correlação de Pearson e Spearman para os dados apresentados na Figura 5 são mostrados na Tabela 7, indicando que não há forte correlação (positiva ou negativa) entre as medidas. Portanto, rejeita-se a hipótese nula H_{2_1}: $\forall c_i \; \exists \mid \text{cor(ACC)} \mid \geq 0{,}7$, em que o tamanho do conjunto de treinamento possui correlação fraca com a acurácia. Mais especificamente, o aumento no tamanho do conjunto de treinamento tem baixo impacto na acurácia do classificador de gênero testado.

7 Aplicando o Classificador de Gênero na Sinalização Digital

O classificador de gênero validado foi aplicado em um cenário real de sinalização digital, em que vídeos do tipo esportivo e jornalístico foram exibidos em uma cantina da Universidade Federal de Campina Grande, Paraíba, enquanto que vídeos dos espectadores assistindo aos anúncios eram capturados. Tais vídeos, foram posteriormente processados com a aplicação do detector de faces e do classificador de gênero. Os vídeos dos espectadores foram capturados em dias alternados e no mesmo período do dia. Os resultados obtidos pelo

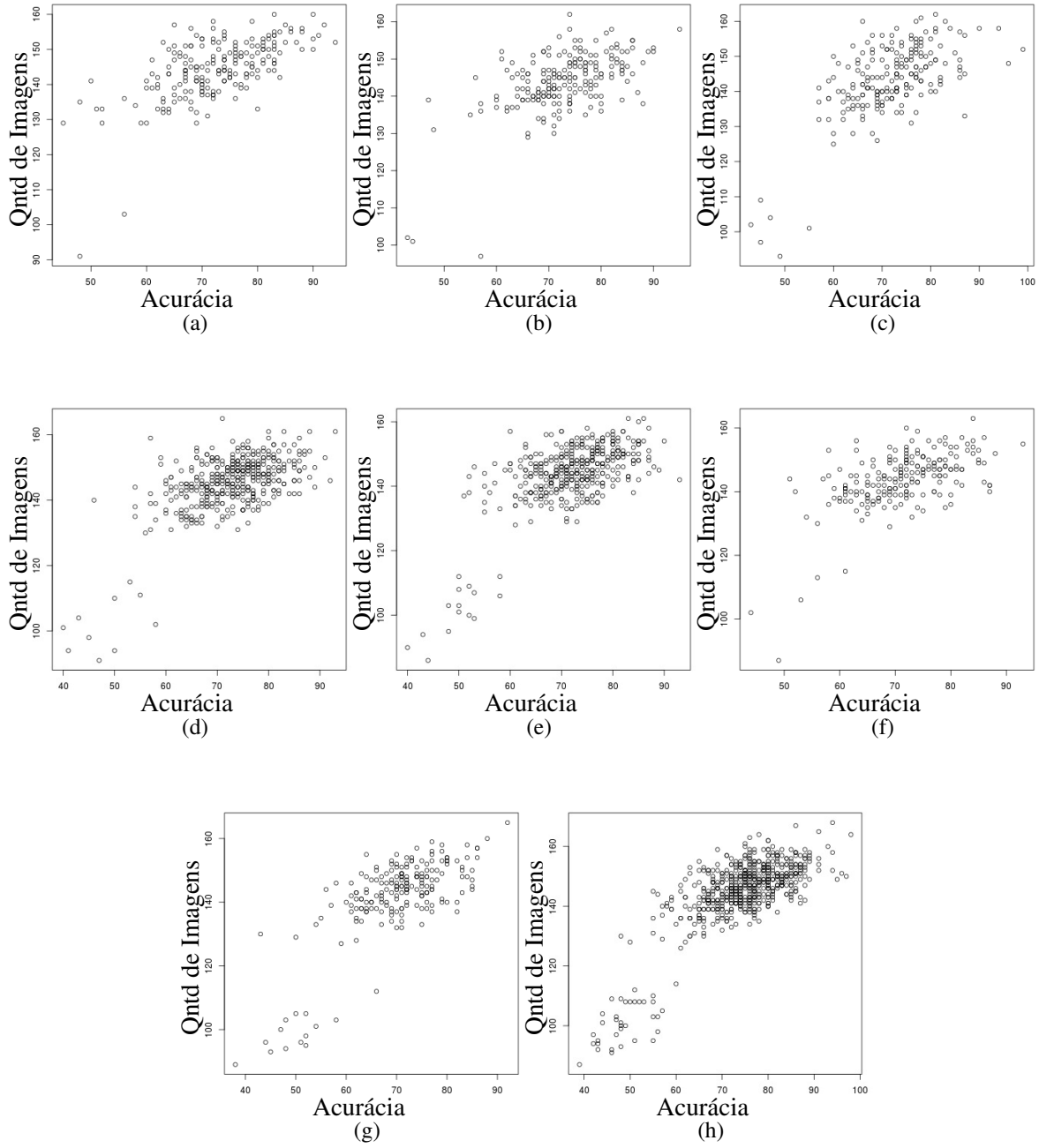


Figura 5. Relação entre a quantidade de imagens no conjunto de treinamento e a acurácia do classificador, utilizando diferentes quantidades de imagens no conjunto de treinamento: (a) 200; (b) 250; (c) 300; (d) 350; (e) 400; (f) 450; (g) 500; (h) 550 imagens, respectivamente.

classificador de gênero nos vídeos coletados são discutidos nesta seção.

Na Figura 6 relaciona-se a quantidade média de pessoas observando a sinalização digital ao tipo de conteúdo que está sendo transmitido, com respectivos Intervalos de Confiança (IC) de 95% no teste t de *student* [3]. Desta maneira, pode-se observar que as pessoas se mantém mais concentradas nas transmissões do tipo esportiva 6 (a), isto é evidenciado com o menor intervalo de confiança nos gráficos, indicando que as pessoas se mantém menos dispersas. O oposto é observado no gráfico das transmissões jornalísticas 6 (b), em que o intervalo de confiança é maior, consequentemente, os espectadores se concentraram menos nestes tipos de anúncios. Tais observações corroboram com as análises feitas por Ravnik e Solina [9], em que o público masculino se concentrou mais nos anúncios do que o público feminino, entretanto, o tipo de conteúdo não foi especificado naquele trabalho.

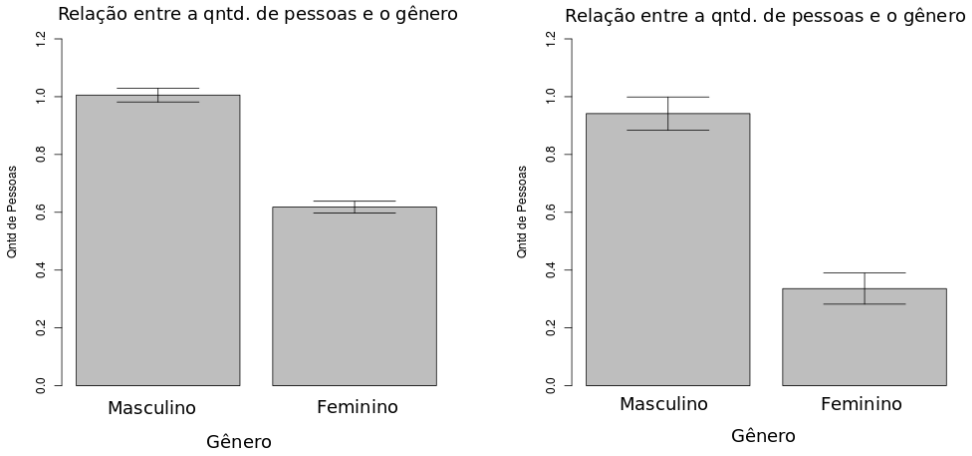


Figura 6. Transmissões de diferentes tipos de vídeo transmitidos: (a) esportivos; (b) jornalísticos

Na Figura 7, pode ser observado que os homens se concentraram mais nos anúncios esportivos do que nos jornalísticos devido ao menor intervalo de confiança [3]. Além disso, o estudo mostrou que homens atendem, de forma estatisticamente igual, aos dois tipos de anúncio. Enquanto isso, as mulheres se concentram mais em anúncios esportivos e, proporcionalmente, menos em anúncios jornalísticos. A partir das observações anteriores, não se pode rejeitar a hipótese nula H_{3_1}: $\mu_M^j = \mu_M^e$, pois há interseção nos intervalos de confiança. Consequentemente, não se pode afirmar a qual conteúdo os homens são mais frequentes. Quanto à hipótese nula H_{4_1}: $\mu_F^j = \mu_F^e$, ela pode ser rejeitada em favor da hipótese alternativa H_{4_2}: $\mu_F^j \neq \mu_F^e$, ou seja, pode-se afirmar que as mulheres ocorrem em maior frequência nos anúncios do tipo esportivo do que nos anúncios do tipo jornalístico. As constatações anteriores se

alinham ao que foi observado previamente em [10] e em [11], em que o gênero do espectador tem uma alta correlação com o conteúdo transmitido.

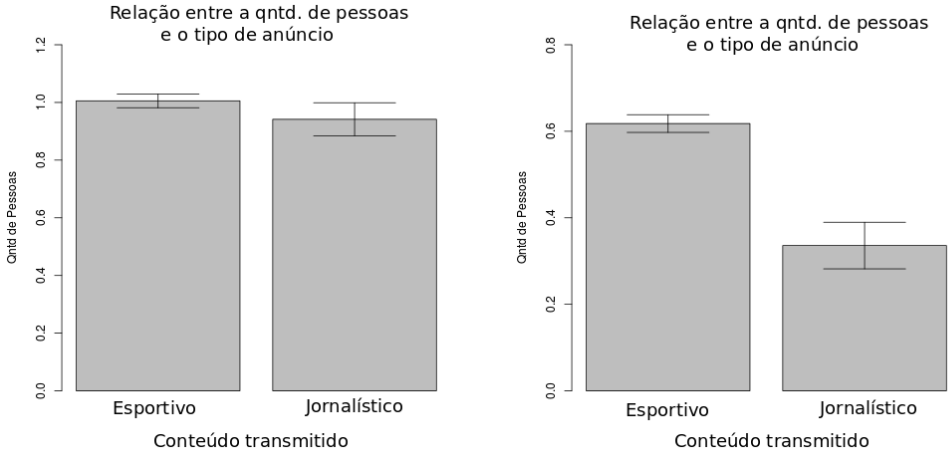


Figura 7. Relação entre o gênero e o tipo de conteúdo: (a) masculino; (b) feminino

Como exemplo de aplicação prática dos achados anteriores, ao observar diferença estatística na atenção dos espectadores considerando seu gênero, o dono de um determinado negócio, pode alterar o conteúdo transmitido automaticamente, à depender do interesse dos clientes, com o propósito de melhorar a eficácia nas transmissões de sinalização digital.

7.1 Empacotamento

Os arquivos utilizados neste experimento foram agrupados para disponibilização pública[8], entre eles, a relação dos vídeos transmitidos no experimento e os *shell scripts* para automatizar a execução do algoritmo de classificação de gênero.

7.2 Ameaças à validade

Como o experimento foi desenvolvido em um ambiente de pequenas dimensões, com circulação limitada de pessoas, os resultados do experimento podem não refletir o comportamento geral de uma ampla população de espectadores. Para reduzir tal ameaça, poderiam ser utilizados outros ambientes mais amplos para replicação e comparação dos resultados. Além disso, a acurácia do classificador de gênero (aproximadamente 80%), pode pode ter impacto na confiança das inferências realizadas.

[8]Endereço para download: https://goo.gl/iS10Hn.

8 Conclusões e trabalhos futuros

O presente trabalho focou na formulação e teste de hipóteses no cenário de sinalização digital. Foram investigadas, preliminarmente, as seguintes relações: quantidade de imagens no conjunto de treinamento versus (tempo de treinamento e acurácia de um classificador de gênero a partir de imagens de faces). Tal classificador foi utilizado em um sistema de análise indireta da atenção dos espectadores, com o propósito de investigar a relação entre o gênero do espectador e conteúdo transmitido.

Foi observado que ao se elevar o número de imagens no conjunto de treinamento, não ocorre aumento significativo na taxa de acerto do classificador, entretanto, como esperado, o tempo de treinamento aumenta consideravelmente. Além disso, foi observado que vídeos esportivos retém maior concentração dos espectadores, independentemente do gênero do espectador, e que homens assistem mais aos anúncios independentemente do conteúdo transmitido, enquanto que as médias de mulheres assistindo anúncios jornalísticos e esportivos são diferentes, i.e., elas estão mais presentes nos anúncios esportivos do que durante os anúncios jornalísticos.

As contribuições centrais deste trabalho residem nas análises realizadas no contexto de sinalização digital, utilizando um modelo de classificação de gênero validado estatisticamente.resultados sem tanta análise estatística, gerando inconsistências em suas análises.

Como trabalhos futuros, pretende-se: (i) aprimorar a acurácia do classificador de gênero para maior confiança nos resultados; (ii) utilizar métodos de rastreamento facial para diminuir o custo computacional do sistema de análise; (iii) considerar outros fatores demográficos, tais como: faixa etária, tempo de visualização de cada indivíduo, etc.; e (iv) incluir exibição de vídeos contendo outros tipos de conteúdo, a fim de avaliar o impacto nas inferências.

9 Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à responsável pelo estabelecimento comercial que cedeu espaço para realização do experimento científico e aos colegas do Laboratório de Percepção Computacional (LPC) que ajudaram com discussões sobre a pesquisa.

Contribuição dos autores:

- Ítalo de Pontes Oliveira: Escreveu primeira versão do artigo, montagem dos equipamentos, coleta dos dados, análise dos resultados.

- Eanes Torres Pereira e Herman Martins Gomes: Orientação na realização do experimento e escrita do primeiro artigo, sugestão de hipóteses, colaboraram na escrita e correções do artigo nas demais versões.

Referências

[1] T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit"criteria based on stochastic processes. *Ann. Math. Statist.*, 23(2):193–212, 06 1952.

[2] Cannes Lions Archive. Strip commerce. http://www.canneslionsarchive.com/entries/528012/strip-commerce. acessado em 3 de Setembro de 2015.

[3] Joan Fisher Box. Guinness, gosset, fisher, and small samples. *Statist. Sci.*, 2(1):45–52, 02 1987.

[4] Intel. Engagin bank customers with interactive digital signage. http://www.intel.co.za/content/dam/www/public/us/en/documents/solution-briefs/engaging-bank-customers-with-interactive-digital-signage-brief.pdf. acessado em Setembro de 2015.

[5] Don-Lin Yang Kuo-Cheng Yin, Hsin-Chieh Wang and Jungpin Wu. A study on the effectiveness of digital signage advertisement. In *2012 International Symposium on Computer, Consumer and Control (IS3C)*, pages 169–172, June 2012.

[6] Alan B. Forsythe Morton B. Brown. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367, 1974.

[7] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895.

[8] POPAI. Screen-media formats. http://www.popai.com/uploads/downloads/POPAIDigitalSignage-Screen-Media-Formats-2009.pdf. acessado em Setembro de 2015.

[9] Robert Ravnik and Franc Solina. Audience measurement of digital signage: Quantitative study in real-world environment using computer vision. *Interacting with Computers*, 25(3):218–228, 2013.

[10] Robert Ravnik and Franc Solina. Audience adaptive digital signage using real-time computer. *International Journal of Advanced Robotic Systems*, 2013-02-07.

[11] Robert Ravnik and Franc Solina. Interactive and audience-adaptive information interfaces. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 13(2), 2014.

[12] Sheeraz Akram Sajid Ali Khan, Muhammad Nazir and Naveed Riaz. Gender classification using image processing techniques: A survey. In *2011 IEEE 14th International Multitopic Conference (INMIC)*, pages 25–30, Dec 2011.

[13] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[14] M. Sharkas and M. A. Elenien. Eigenfaces vs. fisherfaces vs. ica for face recognition; a comparative study. In *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, pages 914–919, Oct 2008.

[15] Potal Digital Signage. Aplicações do digital signage. http://www.portaldigitalsignage.com.br/. acessado em 3 de Setembro de 2015.

[16] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manage.*, 45(4):427–437, July 2009.

[17] C Spearman. The proof and measurement of association between two things. *International Journal of Epidemiology*, 39(5):1137–1150, 2010.

[18] Phil Tian, Addicam V. Sanjay, Kunapareddy Chiranjeevi, and Shahzad Malik Malik. Intelligent advertising framework for digital signage. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'12, pages 1532–1535, New York, NY, USA, 2012. ACM.

[19] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[20] Roy Want and Bill N. Schilit. Interactive digital signage. *Computer*, 45(5):21–24, May 2012.

Autômatos celulares unidimensionais caóticos com borda fixa aplicados à modelagem de um sistema criptográfico para imagens digitais

One-dimensional Chaotic Cellular Automata with fixed border applied to a cryptosystem modeling for digital images

Eduardo C. Silva [1] [2]
Jaqueline A. J. P. Soares [1] [3]
Danielli A. Lima [1] [4]

Data de submissão: 01/03/2016, *Data de aceite: 04/05/2016*

Resumo: O principal objetivo da criptografia de dados é possibilitar que duas entidades se comuniquem ao longo de um canal inseguro, de tal forma que nenhuma outra entidade consiga decifrar a mensagem que é enviada. Muitos métodos de criptografia clássica já foram investigados para minimizar este problema. Uma nova abordagem para este tema são os Autômatos Celulares (ACs), atualmente estudados por sua capacidade de processar grandes volumes de dados em paralelo. Nesse trabalho é investigado um novo modelo de autômato celular para criptografia de imagens, que tem como característica o uso do cálculo de pré-imagens a partir de chaves caóticas. O modelo é denominado Border Chaotic Cellular Automata (BCCA) para cifragem. Resultados mostraram que o modelo tem grande potencial para a realização de cifragem de grandes volumes de dados.

***P*alavras-chave:** autômatos celulares, cálculo de pré-imagens, criptografia de imagens, processamento paralelo

[1]Instituto Federal de Educação Ciência e Tecnologia do Triângulo Mineiro (IFTM) - Av. Líria Terezinha Lassi Capuano nž 255, Bairro Chácara das Rosas, CEP 38740-000, Tel.:(34)3515-2100 - Patrocínio - MG, Brasil.
[2]{eduardocassiano@iftm.edu.br}
[3]{jaquelinepapini@iftm.edu.br}
[4]{danielli@iftm.edu.br}

Abstract: The main objective of data encryption is allowing two entities to communicate over an insecure channel, such that no other entity can decipher the message that is posted. Many classical methods of cryptography have been investigated to minimize this problem. A new approach to this theme are the Cellular Automata (CA), currently studied for their ability to process large volumes of data in parallel. In this work is investigated a new model of cellular automata for encrypting images that uses calculation of pre-images from chaotic keys. The model is denominated Border Chaotic Cellular Automata (BCCA) cipher model. Results showed that the model has great potential for performing encryption of large data volumes.

***K*eywords:** cellular automata , pre-image calculus, image encryption, parallel processing

1 Introdução

Com o aumento dos sistemas de informação conectados à rede mundial de computadores, e posterior popularização de equipamentos eletrônicos para a captura de imagens digitais, tornou-se mais frequente a troca de dados entre entidades, especialmente a troca de imagens em redes sociais privativas ou correios eletrônicos. Além disso, muitas imagens são capturadas a todo instante por satélites artificiais e as imagens obtidas pelos mesmos devem ser transmitidas eletronicamente. Assim, grande parte dos dados que estão sendo transmitidos devem ser protegidos, pois na maioria das vezes esses dados são confidenciais. Adicionalmente, o mecanismo que realiza a proteção desses dados deve ser rápido o suficiente para que esse procedimento seja factível de ser realizado dentro de um tempo de comunicaçao pré-estabelecido. A ferramenta mais significativa para realizar essa tarefa é a criptografia.

O principal objetivo da criptografia de dados é possibilitar que duas entidades se comuniquem ao longo de um canal de transmissão, de tal forma que nenhuma outra entidade consiga decifrar a mensagem que é enviada. A criptografia também é estudada como abordagem para minimização da vulnerabilidade de dados e recursos, bem como para a garantia de confidenciabilidade, integridade e autenticidade durante a transmissão de dados. Esse campo tem sido muito estudado atualmente devido à necessidade constante de transmissão de informações e por serem importantes, na maioria das vezes, devem ter um tratamento especial. Os algoritmos de criptografia clássica existentes não tratam de forma adequada a grande quantidade e redundância de dados, que são características inerentes às imagens. Esse fato deve-se a pouca disponibilidade de núcleos que os dispositivos de baixo custo apresentam. Além disso, os principais algoritmos de criptografia simétrica AES e DES são sequenciais, dificultando o processamento massivo de dados Daemen [4], Zeghid [29]. Isso motivou a criação de um novo campo na criptografia que estuda algoritmos para a cifragem de imagens Yu [28]. Esse campo consiste na melhoria de algoritmos clássicos Prasad [19], na tentativa de paralelizar algumas etapas custosas ou redundantes do processo de cifragem Le [10].

Uma ferramenta útil na elaboração de sistemas de cifragem pode ser estabelecida por modelos matemáticos conhecidos como Autômatos Celulares (ACs). Existem diversas aplicações no uso de autômatos celulares, dentre elas podemos citar, a modelagem de sistemas naturais ou biológicos Lima [11], Zhang [30], físicos Feliciani [6], Castro [3] e até mesmo na proposição de sistemas de controle de robôs Ferreira [7], que seriam muito difíceis de serem modeladas pelas equações diferenciais, sendo estas as mais utilizadas nesse tipo de tarefa Wolfram [26]. Duas das principais vantagens em se aplicar modelos baseados em ACs reside na sua simplicidade de implementação e também na sua capacidade de processar dados em paralelo Vasantha [24].

Neste trabalho, um novo modelo de criptografia para imagens denominado BCCA (Border Chaotic Cellular Automata), baseado em ACs unidimensionais caóticos é investigado. Essa técnica utiliza o cálculo de pré-imagens (evolução do AC para trás) no processo de cifragem, a exemplo de outros métodos investigados na literatura Gutowitz [9], Lima [12] e Oliveira [17, 18, 15]. Para evitar o problema do aumento de bits apresentado no trabalho de Gutowitz [9] e Oliveira [17], neste trabalho, foi utilizada uma borda fixa para bloquear o crescimento de bits acelerado. Além disso, temos a garantia de que todo texto plano pode ser cifrado, diferentemente do trabalho de Oliveira [18], e de que o procedimento aqui empregado é totalmente paralelo, o que não foi observado no trabalho de Oliveira [15]. Adicionalmente, para a comprovação de que o sistema BCCA é forte contra ataques de criptoanálise, testes clássicos em criptografia foram realizados para validar o modelo aqui investigado, tais como, histograma de cores Prasad [19], análise de perturbação e entropia Stinson [23].

2 Fundamentação Teórica

Esta seção apresentará as principais definições para o entendimento do método de criptografia BCCA. Primeiramente, será apresentado uma definição sobre autômatos celulares unidimensionais e a propriedade de regra denominada sensitividade. Na sequência, os principais trabalhos sobre ACs aplicados à criptografia são apresentados e discutidos.

2.1 Autômatos celulares

Um AC é composto por um reticulado com uma dimensão d dividido em células ou unidades processadoras, sendo que, cada célula C é representada por um estado. As células modificam seus estados a cada passo de iteração de acordo com uma regra de transição. Podemos aplicar a regra de transição por T passos de tempo para obter a evolução espaço-temporal do reticulado do AC. A regra estabelecida por uma função de transição indica o novo símbolo a ser escrito na célula do reticulado de acordo com seu estado atual e dos estados de suas vizinhas (regra local). Em sua definição mais usual, a atualização dos estados se dá de forma síncrona e utiliza uma regra determinística, isto é, a cada passo de tempo todas as N

células do reticulado são atualizadas Castro [3].

A estruturação de um AC unidimensional (1D) é a forma mais estudada. Para um AC com regra de atualização determinística, a mudança de estado de uma célula depende de m vizinhas expressas por $m = (2r + 1)$, sendo r o raio do AC Oliveira [16]. Para ilustrar um AC unidimensional com regra de atualização determinista, considere a Figura 1 (a), que aborda uma modelagem conhecida como regra 30 Wolfram [26], contendo um reticulado de 6 células sendo que o estado inicial de cada célula é apresentado em $t = 0$. Uma regra binária de raio 1 é aplicada, sendo que a vizinhança de cada célula é formada por três elementos: a própria célula e suas duas vizinhas adjacentes (à esquerda e à direita). Como esse AC é binário (2 estados possíveis), existem 8 diferentes vizinhanças, da 000 a 111. A regra em si é dada pelos 8 bits de saída associados a cada vizinhança possível: 01111000. Na Figura 1 (b) vemos a atualização do reticulado (evolução para frente) de $N = 6$ por 2 passos de tempo a partir de sua configuração inicial 101110 em $t = 0$. A cada passo, cada célula do reticulado é atualizada identificando-se sua vizinhança e seu novo estado é dado pelo bit de saída correspondente na regra de transição. Observe como exemplo, a célula de símbolo 1, destacada em $t = 0$, seu próximo estado será 0 em $t = 1$. Essa configuração é dada porque a vizinhança 110 de acordo com a regra de transição retorna 0 ($110 \rightarrow 0$), assim a célula central é atualizada para o estado 0. O reticulado é submetido a condições periódicas de contorno, sendo que a primeira célula é vizinha imediata da última, e vice-versa. Aplicando-se esse procedimento para todas as células do reticulado de forma síncrona por T passos de tempo, tem-se a nova configuração do reticulado a cada passo de tempo.

Algumas propriedades estudadas nos ACs são exploradas em criptografia, sendo uma

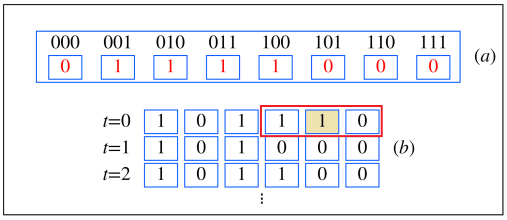


Figura 1. (a) Regra de transição de raio 1. (b) Evolução do AC por *T = 2* passos de tempo.

dessas propriedades a sensitividade da regra. A sensitividade existe em uma regra quando a alteração de um bit extremo da vizinhança provoca necessariamente alteração do bit de saída. Se a alteração for feita no extremo esquerdo da vizinhança e esta provocar alteração nos bits de saída, temos sensitividade à esquerda. Senão temos sensitividade à direita. A regra da Figura 1 (a) apresenta sensitividade à esquerda; por exemplo, a vizinhança **0**00 leva o estado da célula central a 0, enquanto a vizinhança **1**00 leva a 1. Em todos os quatro pares de vizinhanças similares da regra, diferenciadas apenas pelo primeiro bit da regra, a saída é complementar. Dessa forma, a regra da Figura 1 é sensível à esquerda. O método de criptografia investigado neste trabalho, faz parte de uma família de métodos que utilizam

ACs com regras sensíveis e o cálculo de pré-imagem na etapa de cifragem Gutowitz [9] e Oliveira [17, 18, 15] e Lima [12].

2.2 Criptografia Baseada em Autômatos Celulares

O primeiro trabalho conhecido envolvendo o uso de ACs para a realização da criptografia foi proposto por Wolfram [25]. Nesse modelo precursor, a regra de transição é fixa e apresenta dinâmica caótica Wolfram [26]. A chave é utilizada como reticulado inicial a partir do qual a regra é aplicada por um número fixo de passos. Os modelos que utilizam ACs na criptografia com regras sensíveis foi explorado em Sen [22], no entanto o paralelismo e a segurança desses modelos é limitada, devido ao fato da propriedade aditiva das regras, que não as tornam caóticas. Sistemas criptográficos de boa qualidade devem retornar cifras embaralhadas e caóticas Wolfram [26], Machicao [14], dificultando a criptoanálise.

Para melhor compreensão dos algoritmos baseados no cálculo de pré-imagens de ACs algumas definições serão introduzidas. O texto plano $\mathcal{P}$ apresentado neste trabalho refere-se a qualquer sequência de bits dada como entrada, um texto ou uma imagem binarizada. Esse texto plano $\mathcal{P}$ refere-se ao reticulado L inicial do AC. O texto cifrado $\mathcal{C}$ é a sequência de bits final encontrada (borda, bits extras e o próprio reticulado L' final), após o cálculo de pré-imagens baseado em ACs através da chave criptográfica $k \in \mathcal{K}$ por T passos de tempo. O conjunto de chaves criptográficas refere-se às regras de atualização do AC. Neste caso, $\mathcal{K}$ é o conjunto de todas as regras com sensitividade a um dos extremos da vizinhança e caóticas (alta entropia). O processo de decifragem refere-se à evolução para frente do AC sobre $\mathcal{C}$ com a chave k por T passos de tempo.

A criptografia baseada no cálculo de pré-imagens de ACs começou a ser estudada depois que alguns pesquisadores observaram que após as células de um reticulado serem submetidas a T passos de tempo, essas poderiam resultar em uma sequência caótica de configurações. A evolução de um AC pode ser realizada com a aplicação direta da função de transição (evolução para frente) ou a partir do cálculo da pré-imagem (evolução para trás). A evolução para frente (*forward*) é obtida através da configuração inicial do reticulado no instante t, após aplicar a regra de transição por um passo de tempo, obtém uma nova configuração de reticulado no tempo $t + 1$. Esse procedimento pode ser repetido por quantos passos de tempo forem necessários. A evolução para trás (*backward* ou cálculo da pré-imagem) é obtida a partir de uma configuração inicial do reticulado L no instante t, a finalidade é descobrir qual reticulado L' no instante $t - 1$ pode dar origem ao reticulado L no instante t, após aplicar a regra de transição. Também nesse caso, esse procedimento pode ser repetido por T passos de tempo. Se um AC tem exatamente uma pré-imagem para todos os reticulados possíveis, então esse é um AC reversível. A grande vantagem de se utilizar o cálculo de pré-imagens baseado em ACs reside no fato de que o procedimento é paralelizável, quando existe uma unidade de processamento paralela para o cálculo de cada célula Anghelescu [1].

Um dos primeiros modelos que empregam o cálculo de pré-imagem com regras sensí-

veis e com dinâmica caótica foi proposto em Gutowitz [9]. Nesse modelo, a regra de transição corresponde à chave criptográfica, o texto original define o reticulado inicial e o cálculo de pré-imagens consecutivas corresponde à etapa de cifragem. Enquanto a decifragem é feita utilizando-se a evolução tradicional do AC (forward). Para a garantia de existência de pré-imagem para qualquer reticulado, regras com a propriedade de sensitividade são empregadas. Entretanto, esse método gera um texto cifrado maior que o texto plano. O método de cifragem proposto por Gutowitz [9] será detalhado e o mesmo está ilustrado na Figura 2. Para demonstrar esse método, será utilizada a regra 30 (chave) com sensitividade à esquerda. A cada passo de tempo, será necessário iniciar os $m-1$ bits à direita do novo reticulado (representado na Figura 2 como células de coloração azul). Uma vez que os bits tenham sido inicializados, começa a etapa de evolução dos bits através da regra de transição juntamente ao valor da vizinhança gerada, preenchendo de forma síncrona e paralela todas as demais células posicionadas à esquerda no reticulado. Além disso, pode-se iniciar paralelamente o preenchimento de instantes de tempo $t+i$, com $i>0$, desde que os bits dependentes de $t+i$ já tenham sido calculados. Esse procedimento acelera muito o processo, pois cada célula é calculada de forma independente das demais. O preenchimento baseia-se no seguinte processo: dada uma vizinhança $?01 \rightarrow 1$ ($?$ é o bit vermelho, 01 são os bits em azul e $\rightarrow 1$ é o bit verde) no passo $t=1$, anexa-se na célula adjacente à cadeia já calculada, o bit de vizinhança extremo à esquerda que provoque como saída o valor 1 em $t=0$. Logo, o valor encontrado no exemplo será 0, pois a regra $001 \rightarrow 1$ (isto é, 001 é a vizinhança que gera o valor de saída 1). O cálculo de pré-imagem será executado até alcançar o número de passos de tempo T estipulado. Uma desvantagem do modelo é a propagação da perturbação, que neste caso aconteceu apenas do lado da sensitividade da regra. Sabe-se que um bom método criptográfico deve propagar a perturbação ao longo de todo o reticulado.

Posteriormente, em Oliveira [17] também é apresentado um modelo que aumenta o

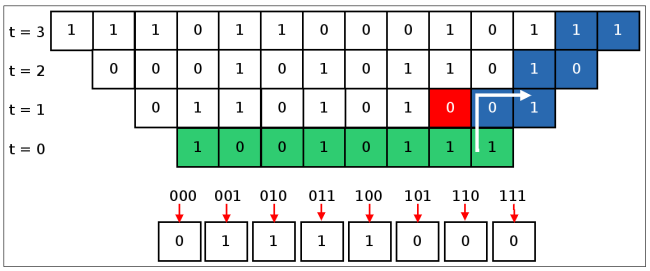


Figura 2. Metodologia aplicada à cifragem de regra sensível à esquerda no modelo de Gutowitz [9].

tamanho do reticulado. No entanto, neste modelo a propagação da perturbação dos bits é propagada ao longo de todo o reticulado por utilizar regras com sensitividade bidirecional, diferentemente do modelo de Gutowitz [9]. Em Lima [12] e Oliveira [17], uma abordagem que preserva o tamanho do reticulado, tornando o texto cifrado do mesmo tamanho que o

original, foi investigada a partir do algoritmo proposto em Wuensche [27]. Contudo, ela tem a desvantagem de que nem todo texto fornecido como entrada pode ser criptografado. No trabalho de Oliveira [15], uma abordagem que faz uso do aumento do texto plano apenas quando é necessário foi elaborado. No entanto, esse modelo apresenta a desvantagem do paralelismo ser quebrado devido à utilização de pilhas de armazenamento global e local que guarda as possíveis pré-imagens. Outros modelos foram posteriormente estudados nessa temática Macedo [13], Barros [2], no entanto, não são estratégias puramente baseadas no cálculo de pré-imagens altamente paralelizáveis de ACs.

No modelo apresentado neste trabalho, a adição de bits extras se faz necessária, no entanto, a quantidade necessária é menor que nos trabalhos de Gutowitz [9] e Oliveira [17], e todo texto de entrada retorna uma pré-imagem válida, o que não acontece no trabalho de Oliveira [18]. Além disso, o método BCCA proposto neste trabalho é altamente paralelizável, diferentemente do modelo de de controle por memórias do tipo pilha em Oliveira [15].

3 Modelo Proposto

Nesta seção serão apresentadas duas abordagens para a cifragem de imagens a partir do método BCCA. A primeira refere-se ao modelo geral de cifragem e aplica-se diretamente às imagens preto e branco no formato *.pbm* (Portable Bitmap Format), onde cada pixel (bit) representa uma cor diretamente. Ou seja, a entrada é uma sequência de bits, onde o bit 0 representa a cor branca e o bit 1 representa a cor preta. Em seguida uma abordagem para a quebra de blocos e embaralhamento em imagens coloridas no formato *.ppm* (Portable Pixmap Format) no padrão RGB é utilizado. Essa quebra em blocos se faz necessária porque as imagens nesse formato apresentam o sistema de cores no padrão RGB formando-se blocos muito grandes. Após o embaralhamento e transformação em blocos menores de bits, a cifragem é realizada através do algoritmo BCCA.

3.1 Abordagem para cifragem de imagens preto e branco

O método de cifragem BCCA é um método para cifragem de blocos unidimensionais de bits. Uma imagem preto e branco é uma matriz de bits, onde cada linha é uma sequência de bits. O BCCA consiste na evolução para trás (*backward*) do AC e é obtido a partir de uma configuração inicial do reticulado L, que é considerado o texto plano $\mathcal{P}$, no instante t, a finalidade é descobrir qual reticulado L' no instante $t-1$ pode dar origem ao reticulado L no instante t, após aplicar a regra de transição, que representa a chave criptográfica. Esse procedimento pode ser repetido por T passos de tempo resultando num padrão caótico (texto cifrado $\mathcal{C}$). Para evitar que o texto cifrado aumente de tamanho Gutowitz [9], neste trabalho foi proposta a criação de uma borda fixa, que é a responsável por limitar esse crescimento de bits. Qualquer sequência pseudo-aleatória pode ser utilizada na composição dessa borda,

inclusive a sequência proposta por Wolfram [26]. No entanto, neste trabalho, a borda refere-se à sequência de bits que corresponde a metade da chave criptográfica k de $r = 1$. A Figura 3 apresenta a chave criptográfica dada por $\{01111000\}$ e a borda b em azul escuro $\{1000\}$, que é a metade dessa chave criptográfica. Se mais bits precisarem ser enviados, eles serão concatenados à sequência de bits da borda b. O próximo bit a ser concatenado é o bit 0 e a borda torna-se 01000. A borda deve ser armazenada e enviada junto ao texto cifrado.

O processo de cifragem de uma linha $L = \{10010111\}$ da imagem preto e branco com $N = 8$ pixels está apresentado na Figura 3. Para detalhar o procedimento, um exemplo será apresentado para a atualização do bit em vermelho no tempo $t = 1$, considerando-se o preenchimento do bit ?, de acordo com a seguinte vizinhança $?01 \rightarrow 0$. A partir da regra de transição utilizada como chave criptográfica, temos que $101 \rightarrow 0$. Assim, o primeiro bit é atualizado e todos os demais bits também o são, partindo-se da mesma definição. Para a aplicação desse método em imagens, essas devem ser quebradas em blocos unidimensionais. O blocos são lineares e representam uma linha ou uma coluna inteira de bits (pixels) da imagem. Assim, o método pode ser aplicado em cada bloco da imagem por T passos de tempo. O custo de processamento P_{CB} para a realização da cifragem de um bloco de tamanho N da imagem é realizado em $P_{CB} = T + N + 2 \times (r - 1)$ ciclos de *clock*, sendo que $N = \max(m, n))$, tal que, $N = n$. No exemplo da Figura 3 temos que $P_{CB} = 3 + N + 2 \times (1 - 1) = 11$ ciclos de *clock* se existirem T núcleos de processamento (3 núcleos). A cifragem de uma imagem de tamanho $n \times m$ é realizada com um custo de processamento de $P_{CI} = m \times (T + N + 2 \times (r - 1))$ ciclos de *clock* do computador.

Se uma cifragem a partir do AES de uma imagem com as mesmas dimensões for comparada com o modelo proposto neste trabalho, o tempo medido em ciclos de *clock* seria de $P_{CB} = (T \times (2 \times N + N^2 + N^3))$ para um bloco de linha. Para o tempo de cálculo de uma imagem teríamos, $P_{CI} = m \times (T \times (2 \times N + N^2 + N^3))$, tal que T é o número de *rounds* necessários para cifrar um bloco de tamanho N bits Daemen e Rijmen [4] - neste caso, uma linha da imagem. Assim, temos que para a cifragem da linha da Figura 3 tem-se que $P_{CB} = 3 \times (2 \times 8 + 8^2 + 8^3) = 3288$ ciclos de *clock*. Esse atraso deve-se ao fato de que todos os blocos são concatenados através de uma função $\odot$ XOR.

O método de decifragem, adotado neste trabalho, baseia-se na evolução de um AC e é realizada com a aplicação direta da função de transição na evolução para frente (*forward*) do AC junto ao texto cifrado e a borda (armazenada), por T passos de tempo. Dessa maneira, o texto plano é obtido novamente. O custo de processamento P_{DB} para a realização da decifragem de um bloco de tamanho N da imagem é realizado em $P_{DB} = T$ ciclos de *clock*, sendo que $N = \max(m, n))$ e vamos supor que $N = n$. No exemplo da Figura 3 temos que a linha de tamanho N seria decifrada em $P_{DB} = 3$, se existirem 8 núcleos de processamento - um para cada célula do reticulado N. A decifragem completa de uma imagem de tamanho $n \times m$ é realizada com um custo de processamento de $P_{DI} = m \times T$ ciclos de *clock* do computador. Para efeito de comparação, o algoritmo AES levaria a mesma quantidade de tempo da cifragem para decifrar uma imagem de tamanho, ou seja, $P_{DI} =$

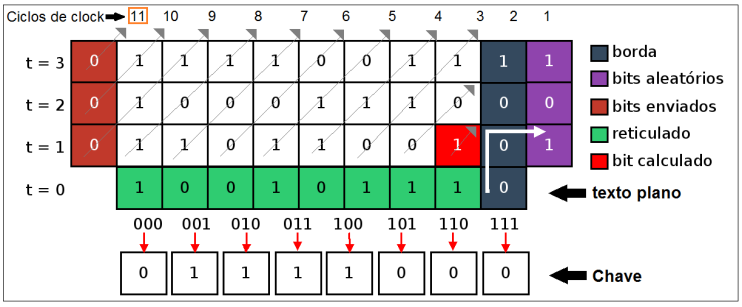


Figura 3. Exemplificação do processo de cifragem do método proposto.

$m \times (T \times (2 \times N + N^2 + N^3))$ ciclos de *clock*. Assim, o algoritmo AES levaria para cifrar uma linha $P_{CB} = 3288$ ciclos de *clock*. Portanto, a investigação da abordagem paralela baseada em ACs faz-se necessária, uma vez que reduz significativamente o tempo de cifragem e decifragem, quando implementado num hardware paralelo. Isso significa que se existir um hardware capaz de processar cada célula do AC ele vai levar menos tempo para cifrar que o algoritmo AES, uma vez que o algoritmo AES possui diversas partes (rounds) que são dependentes entre si e sequenciais Daemen e Rijmen [4].

3.2 Abordagem para a cifragem de imagens coloridas

A aplicação de uma etapa de pré-processamento no modelo para as imagens coloridas *.ppm* antes da aplicação do modelo de criptografia BCCA aqui proposto, se faz necessária, uma vez que a codificação empregada em tais arquivos é mais complexa do que imagens em preto e branco do tipo *.ppm*. Esse fato é verificado uma vez que o algoritmo BCCA, proposto neste trabalho é capaz de cifrar apenas sequências de bits. Lembrando que qualquer sequência de bits poderia ser cifrada através do algoritmo proposto, incluindo-se arquivos de voz, de vídeos ou de texto. Como o objeto de estudo para validação do algoritmo BCCA escolhido foi as imagens digitais coloridas, então, uma pré-processamento foi necessário para atingir uma cifra segura e sem texturas na imagem cifrada, que a qualidade da cifragem.

A principal diferença entre uma imagem *.pbm* e uma imagem *.ppm*, deve-se ao fato que uma imagem colorida possui mais de um único descritor de cor para cada pixel. O esquema de cores utilizado neste trabalho é constituído do padrão RGB (*red*, *green* e *blue*), isto é, a união de três descritores de cor que variam seus valores entre 0 e 255, originando 16.777.216 possibilidades de cor para um único pixel. Utilizando-se do RGB é possível reproduzir figuras possuindo paletas de cores complexas.

Comparada à versão preto e branco, a implementação do novo modelo difere apenas na etapa de formação dos blocos para cifragem. Uma vez que um novo bloco é criado, a cifragem e decifragem é realizada da mesma maneira, ou seja, utilizando-se o algoritmo BCCA. De acordo com o novo modelo as figuras coloridas empregadas no processo recebem um

pré-embaralhamento nos descritores de cor RGB, presentes nos pixels da imagem. A técnica consiste em separar os bits de cada descritor e realizar concatenações formando novos blocos destinados à encriptação. Realizar esse procedimento eleva o nível de segurança do modelo, já que é proporcionado inicialmente um certo nível de embaralhamento, evitando zonas de texturas Prasad [19]. No processo de decifragem esse embaralhamento é realizado da mesma maneira, recuperando-se cada um dos blocos de cor novamente.

A Figura 4 representa um exemplo sobre as primeiras etapas do embaralhamento dos canais RGB. Primeiramente, divide-se a imagem original representada na Figura 4 (a) em 4 partes distintas ilustradas na Figura 4 (b). Em seguida, o procedimento extrai os canais RGB de cada um dos blocos, separando-os por tipo e consequentemente formando seções que possuem todos os bits identificadores das cores. A Figura 4 (a) apresenta o resultado do primeiro bloco da Figura 4 (c), que resultou em 3 blocos de matrizes de $32 \times 32 \times 24$ bits. Ou seja, 24 matrizes de 32×32 bits cada. Dessa forma, são formados 12 blocos, contendo todos os descritores de cor contidos na imagem.

As próximas etapas caracterizadas pela Figura 5, demonstram a fase de formação

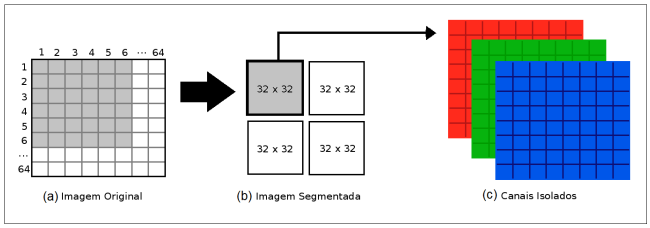


Figura 4. Primeiros passos da quebra em blocos na cifragem de imagens coloridas.

dos blocos finais que serão utilizados na cifragem. A Figura 5 (a) representa o bloco de $32 \times 32 \times 24$ bits da Figura 4 (b). Inicialmente, cada coluna de cada uma das 24 matrizes de bits de 32×32 bits são concatenadas, conforme é apresentado na Figura 5 (b). O novo bloco resultante pronto para a aplicação do modelo BCCA possui 24×32 bits. Posteriormente, o procedimento é repetido para as demais colunas, conforme apresentado na Figura 5 (c). Por fim, o procedimento é repetido para a última coluna, conforme é apresentado na Figura 5 (d). Assim, tem-se 32 blocos de 24×32 que serão cifrados utilizando-se o algoritmo BCCA.

A Figura 6 apresenta um exemplo mais detalhado sobre o embaralhamento. A imagem 2×2 é segmentada em 4 partes, contendo cada uma, 3 blocos descritores de cor. Inicialmente, cada bloco apresenta o tamanho de $1 \times 1 \times 24$ bits, que servirá na formação de uma linha que será utilizado como bloco de cifragem. Na sequencia, cada descritor de cor de tamanho 8 será dividido, formando para cada pixel $3 \times (8 \times 1 \times 1)$ bits. Em seguida cada um desses bits serão concatenados à formando um bloco de $1 \times 1 \times 24$, com os bits alternados representando cada um dos descritores de canal RGB. Para detalhamento da formação do bloco do pixel vermelho, tem-se o primeiro bit do bloco representado pelo primeiro bit do canal R. O segundo bit do bloco será representado pelo primeiro pixel do bloco G. Posteriormente, o

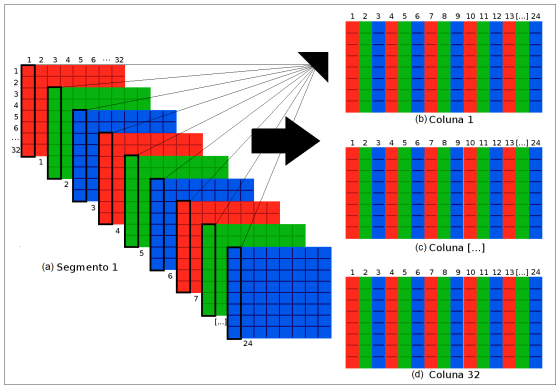


Figura 5. Método aplicado à composição dos blocos de cifragem para imagens coloridas.

terceiro bit do bloco de cifragem será o primeiro pixel do bloco B. Assim, todos os demais bits são concatenados ao bloco (linha) que será cifrado para o pixel vermelho. Em seguida, o procedimento é repetido para os demais pixels da imagem 2×2, de cores verde, azul e branco. Assim, tem-se um bloco de tamanho 4×24 bits para ser cifrado por linhas através do algoritmo BCCA.

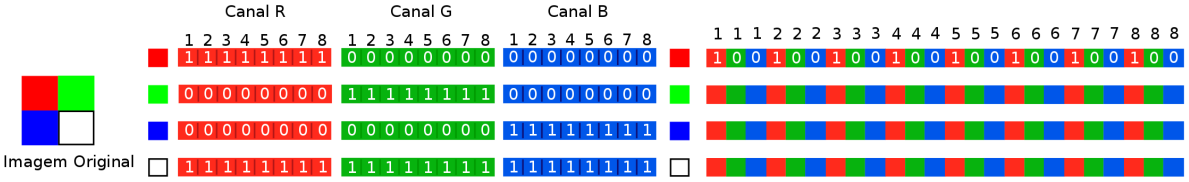


Figura 6. Exemplo de aplicação utilizando o embaralhamento proposto neste trabalho para a cifragem de imagens coloridas.

4 Metodologia para validação de resultados

Algumas das definições apresentadas a seguir serão importantes para a compreensão dos experimentos realizados neste trabalho para demonstrar que o modelo de criptografia baseado em ACs unidimensionais caóticos BCCA aqui apresentado é seguro à ataques de oponentes. Os métodos aqui apresentados fazem parte dos métodos clássicos para validação de modelos criptográficos, dentre eles, entropia, análise de propagação da perturbação e histograma de cores. Toda a implementação dos métodos de criptografia foi realizada de maneira sequencial, uma vez que neste trabalho queremos apenas testar a qualidade de cifragem final. Uma implementação paralela, apenas reduziria o tempo da cifragem, mas não o

resultado final da qualidade da imagem cifrada. Assim, testes de tempo de uma implementação sequencial foram realizados apenas para contrastar o tempo de cifragem de uma imagem colorida e de uma imagem preto e branco com a mesma quantidade de pixels.

4.1 Entropia

Definir através de uma operação matemática a capacidade de confusão presente em uma regra de transição é a principal tarefa do cálculo da entropia. O procedimento baseia-se no mapeamento de janelas contidas em uma determinada cadeia binária (neste caso, a regra de transição). Estas, são abstrações que representam através de seus segmentos cada um dos possíveis valores presentes na cadeia analisada. Com o levantamento da quantidade de manifestações de cada uma das janelas, é possível computar um valor de maneira a definir o grau de confusão (caótica) contido nas evoluções de um AC em conjunto com a regra analisada.

Por exemplo, dada a cadeia binária 01111000 correspondente à regra 30, o tamanho total das possíveis janelas j será obtido por $j = log_2 8$. Em seguida, será executado o mapeamento em busca da quantidade de ocorrências de cada janela. Nesta etapa a regra assume uma composição ligando início e fim, caracterizando uma estrutura em anel, que em função do valor de j, alguns elementos da cadeia deverão se conectar a sua outra extremidade para completar o conjunto imposto, consequentemente alcançando todas as ocorrências possíveis. A Figura 6.5 ilustra o mapeamento com ênfase na busca das janelas analisando o penúltimo e o último elemento, Figuras 7 (a) e 7 (b), respectivamente.

O valor da entropia é dada por $S = \sum_{j=1}^{2^j} \frac{k}{2^j} log_2 \frac{k}{2^j}$, com j sendo o tamanho da

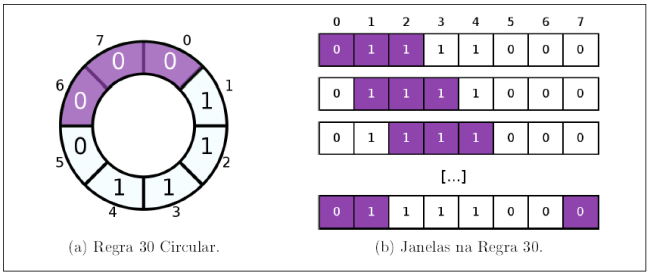


Figura 7. Mapeamento da quantidade de janelas na regra 30 calculado através da entropia s para a verificação da confusão de uma chave criptográfica.

janela e k representa a quantidade de vezes que cada janela aparece na regra. Assim, para o exemplo da Figura 7 temos que $S = 2.5$, pois a janela 000 e 111 aparece 2 vezes, as janelas 001, 011, 100 e 110 aparecem 1 vez e as janelas 010 e 101 não apresentam nenhuma ocorrência. Para promover a normatização dos valores entre 0 e 1, divide-se o valor obtido em $S = 2.5$ pela delimitador da dimensão das janelas ($j = 3$), obtendo o valor de entropia $s = 0.8333$. Quanto mais próximo de 1 estiver o valor de s, maior será o grau de confusão

presente na regra de transição.

Definir a entropia de uma chave é um importante classificador utilizado neste trabalho, uma vez que regras com alta entropia apresentam maior grau de embaralhamento. Logo, regras que possuem baixa entropia não serão interessantes no uso da criptografia baseada em ACs e posteriormente, não farão parte dos experimentos futuros. Dessa forma, o modelo contemplará um conjunto de chaves pré-filtradas, com padrões com maior entropia, onde espera-se obter resultados mais significativos.

4.2 Análise de propagação da perturbação

O nível de embaralhamento ótimo do texto final cifrado é aquele que apresenta média $\bar{x} = 0.5$ de bits 1 (ou bits 0). Ou seja, num texto cifrado ótimo é necessário que se metade dos pixels seja formada por 0s e a outra metade seja formada por 1s. No entanto, essa análise não é suficiente porque os 0s e 1s podem estar agrupados em uma porção da imagem. Assim, a análise da propagação da perturbação se faz necessária. O diagrama ilustrado na Figura 8 representa o experimento da propagação da perturbação. A base do teste é formada por

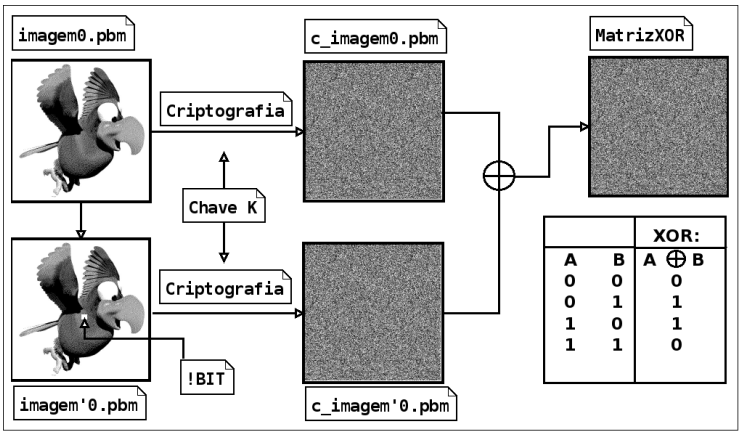


Figura 8. Método de teste denominado propagação da perturbação.

duas imagens idênticas, exceto pela segunda possuir o valor de seu bit central alterado. Em seguida, ambas as figuras passam pelo processo criptográfico empregando a mesma chave e formando respectivamente seus textos cifrados. Posteriormente o experimento utiliza o operador lógico XOR entre as cifras geradas, originando uma matriz que contempla a análise visual e comparativa entre as cifras. A comparação efetuada pelo XOR consegue identificar bits que sofreram modificações no processo de cifragem. Para a avaliação foi utilizado a contagem de 0s na imagem resultante após a aplicação da função XOR. Um cifra de boa qualidade é aquela que apresenta média de 0.5 de 0s e 1s.

4.3 Histograma de cores

Um histograma de uma imagem digital colorida é um gráfico que representa os valores das tonalidades de cores dos pixels nessa imagem. O histograma mostra o número de pixels da imagem com uma cor particular e representa os números no gráfico. Um histograma RGB mostra os três canais combinados, mas um canal de cor individual pode confirmar quais as cores específicas que apresentam picos numa imagem. A escala de cor estende-se ao longo do eixo x horizontal e vai de 0 a 255, a gama da escala é relativa a uma profundidade de cor de 8 bits para cada um dos canais de cores do padrão RGB. O eixo y vertical é o número de pixels contidos numa tonalidade particular. Quanto maior for o valor vertical, mais pixels têm uma tonalidade particular. Em criptografia, dada uma imagem num padrão de cores qualquer, é esperado que ao longo do processo de cifragem seja obtido um histograma de canais individuais equilibrados. Ou seja, se todos os três canais (vermelho, verde e azul) mostrarem gráficos com picos no mesmo ponto com os canais de cores individuais, isso indica que o equilíbrio de cores está corretamente definido e que o algoritmo foi capaz de misturar o padrão de cores da imagem original. Por outro lado, se existir uma variação significativa entre os canais, então é necessário proceder a algum ajustamento para obter o equilíbrio correto.

5 Experimentos e análise de resultados

Esta seção apresenta os experimentos realizados para as imagens preto e branco e coloridas através da cifragem de imagens de diversos tamanhos pelo algoritmo BCCA. As imagens coloridas recebem um tratamento especial, denominado neste artigo de pré-embaralhamento. Além disso, um estudo sobre a qualidade das chaves (regras) é apresentado através do estudo da entropia em Oliveira [15]. Adicionalmente, um estudo elaborado para a qualidade das imagens cifradas é realizado através da metodologia denominada análise da propagação da perturbação. As imagens coloridas também foram avaliadas através do histograma de cores, uma métrica muito importante na definição da segurança dos métodos de criptografia. Por fim, uma análise do tempo de cifragem é analisado para verificar o quanto os parâmetros do modelo afetam o tempo de processamento.

Além disso, é importante considerar que o algoritmo proposto foi implementado na Linguagem C padrão de programação de maneira sequencial, e o Sistema Operacional utilizado foi Debian 7.8 Wheezy, com um Kernel $x86\_64$ Linux 3.2.0-4-amd64 com Processador Intel Core i3 CPU M 370 @ 2.399GHz e Memória RAM 3765 MB. Essa configuração de máquina foi utilizada em todos os experimentos apresentados neste trabalho.

5.1 Análise da quantidade de bits enviados

O primeiro experimento refere-se à análise da quantidade de bits enviados. Considerando-se cada pixel como um bit, os gráficos da Figura 9 (a) e (b) demonstram a quantidade de bits salvos somando os bits originais da imagem 64×64, variando-se os tamanhos de raio (tamanho da chave) por $T = 64$ passos de tempo. Os gráficos da Figuras 10 (a) e (b) apresentam o comportamento da função para cada raio (tamanho de chave) para o modelo de Gutowitz [9] e no proposto, respectivamente. A função que representa esse crescimento no envio de bits é dada por $(m \times n) + (2r \times T \times N)$, para o modelo de Gutowitz [9], sendo que $N = \max(m, n)$. No BCCA esse envio de bits é inferior e é dado por $(m \times n) + (r \times T \times N)$. A quantidade de espaço que precisa para armazenar os bits extras no algoritmo proposto é duas vezes menor que no modelo precursor proposto por Gutowitz [9] e Oliveira [17]. Apesar do novo modelo reter parte dos bits envolvidos no processo, a quantidade total é inferior à apresentada modelo de Gutowitz [9]. Conclui-se que a melhoria proposta neste estudo interrompe a expansão do reticulado no processo criptográfico e reduz a quantidade de bits salvos no texto cifrado.

A Figura 11 apresenta uma comparação entre os modelos de Gutowitz [9] juntamente

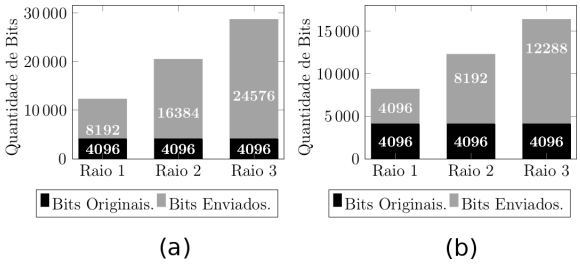


Figura 9. Número de bits enviados: (a) no modelo de Gutowitz [9] e Oliveira [17], e (b) no modelo proposto BCCA.

ao apresentado neste trabalho através da cifragem por $T = 1$ passos de tempo de uma imagem de 16×16 pixels. A Figura 11 (b) e a Figura 11 (c) representam, respectivamente, o modelo Gutowitz [9] e o algoritmo proposto. Conforme foi dito anteriormente, a imagem cifrada pelo algoritmo proposto mostra-se mais reduzida em relação ao método precursor de Gutowitz [9].

5.2 Definição de chaves para a boa qualidade do método de cifragem proposto

O segundo experimento realizado refere-se à análise da qualidade do método para a cifragem de 106 imagens preto e branco de 64×64 pixels no formato *.pbm* (Portable Bitmap Format), que corresponde à 64×64 bits. Para apresentar cifras seguras contra ataques de

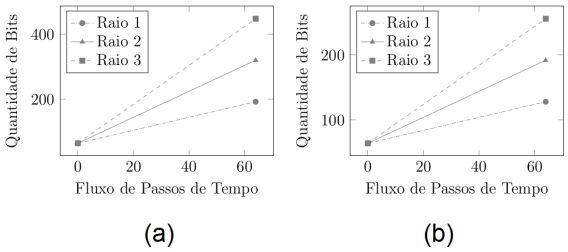


Figura 10. Gráficos das funções de crescimento do número de bits: (a) no modelo de Gutowitz [9] e Oliveira [17], e (b) no modelo proposto BCCA.

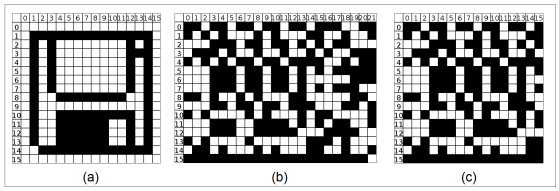


Figura 11. (a) Imagem original. (b) Imagem criptografada resultante a partir do modelo de [9]. (c) Imagem criptografada resultante a partir do modelo proposto neste trabalho.

criptoanálise, é necessário avaliar o embaralhamento do texto cifrado em relação à qualidade das chaves utilizadas na cifragem. Para que um método de criptografia seja seguro, é necessário que o mesmo apresente média de $\bar{x} = 0.5$ de 0s a partir da **análise de propagação de perturbação** entre a cifragem de duas imagens. A primeira imagem representa a imagem original A e a segunda imagem A' se diferencia da primeira através de um bit. Assim, bons métodos de cifragem possuem a mesma quantidade de 0s e 1s na avaliação dessa imagem XOR resultante.

Utilizando-se a entropia s em Oliveira [15], foi possível classificar as 2^{32} chaves de 32 bits que são ao mesmo tempo sensíveis à esquerda Gutowitz [9] e possuem $s \geq 0.7$, resultando em 63872 chaves que satisfazem a esses dois requisitos. Regras com $s = 1.0$ apresentam um nível maior de embaralhamento das chaves criptográficas e portanto tendem a apresentar resultados mais satisfatórios. Em Oliveira [15] foi mostrado que chaves com entropia inferior à $s < 0.7$ não apresentam bom desempenho na cifragem de imagens. A Tabela 1 exibe o resultado referente à cifragem das 106 imagens de 64×64 bits cifradas pelas 63872 regras de raio 2 (chaves de 32 bits) com sensitividade à esquerda, tal que, $0.7 \leq s \leq 1.0$. A faixa de valores da entropia foi dividida em classes para melhor compreensão dos resultados. A média de 0s após a aplicação da **análise da perturbação** está próximo à $\bar{x} = 0.5$ em todas as análises de classes de regras, o que indica que o método pode ser avaliado como caótico, justamente, o que é esperado em algoritmos de criptografia. Além disso, foi mostrado que o

desvio padrão em todos os resultados é inferior à 0.05 em todos os casos, exceto para chaves com entropia mais baixa (Classe 1), onde o resultado do desvio padrão em relação à média dos experimentos foi de 0.1. A variância segue esta mesma análise, ou seja, quanto menor o valor da entropia, maior é a variância em relação à média encontrada e à medida que o valor da entropia aumenta, essa variância diminui. A Tabela 1 também apresenta a variância em relação ao valor ótimo desejado (0.5 de 0s) e o desvio padrão em relação ao valor ótimo é apresentado e mais uma vez, mostra-se que regras com entropia mais baixas (Classe 1) tem dificuldades para cifrar textos planos. Em conclusão aos experimentos das regras sensíveis à esquerda tem-se que o experimento apresentou um bom resultado médio entre todas as classes. Outro fato relevante deve-se à alta entropia da classe 5, alcançando os melhores resultados médios dentre os presentes.

Na segunda fase de experimentos foi possível classificar as 2^{32} chaves de 32 bits que

Sensitividade	Classe	Entropia	Média	Variância	Desvio P.	Var. Ótima	Desv. P. Ótimo
Esquerda	1	0.70 até 0.74	0.542760	0.012041	0.109731	0.013869	0.117768
	2	0.74 até 0.79	0.514253	0.003049	0.055215	0.003252	0.057025
	3	0.80 até 0.84	0.516919	0.003017	0.054931	0.003304	0.057478
	4	0.85 até 0.89	0.511432	0.001475	0.038408	0.001606	0.040073
	5	0.90 até 1.00	0.509390	0.000890	0.029836	0.000978	0.031279
Direita	1	0.70 até 0.74	0.508656	0.000790	0.028100	0.000865	0.029403
	2	0.74 até 0.79	0.505782	0.000368	0.019178	0.000401	0.020031

Tabela 1. Propagação da perturbação em raio 2 e sensitividade à esquerda e à direita.

são ao mesmo tempo sensíveis à direita e possuem $s \geq 0.7$, resultando em 40208 chaves que satisfazem a esses dois requisitos. E essas regras foram classificadas nas Classes 1 e 2 da Tabela 1 e cifraram cada uma das 106 imagens de 64×64 pixels. As análises que podem ser extraídas dos experimentos referem-se também que as regras da Classe 2 realizam cifragens melhores que as regras da Classe 1. Tanto na abordagem com regras de sensitividade à esquerda e à direita os resultados se mostraram de boa qualidade frente ao teste de análise da propagação da perturbação para o algoritmo de criptografia baseado em ACs proposto neste trabalho.

5.3 Análise do tempo de processamento de imagens preto e branco

A implementação empregada neste experimento foi a abordagem sequencial do algoritmo BCCA. Esta implementação tem como objetivo comparar as diferenças no tempo de cifragem em relação à diferenças de parâmetros no próprio modelo. Além disso, a implementação sequencial realizada não tem como objetivo comparar o tempo com outros métodos com abordagem sequencial, tais como, DES e AES. Pois sabe-se que toda a motivação para o emprego dos algoritmos baseados em ACs vislumbram uma implementação paralela Wolfram [26]. Para isso, a implementação deve ser realizada em um hardware de altíssimo desempenho, tais como placas FPGA Das [5]. No entanto, sabe-se que tanto a implementação

paralela quanto a sequencial resultam a mesma imagem cifrada. Portanto, a implementação sequencial pode ser aplicada quando o estudo é apenas para analisar a qualidade da cifragem ou quando não se tem disponível a quantidade de núcleos gasta para a realização do experimento paralelo.

O experimento desta seção resume-se em calcular a quantidade de tempo utilizado na encriptação ou decriptação de diversas dimensões de uma mesma imagem. Para isso, varia-se a quantidade de passos de tempo (isto é, para cada dimensão, o teste avaliará o tempo necessário para cifrar e decifrar uma determinada figura). A regra empregada ao experimento possui raio 2 e é sensitiva à esquerda.

Inicialmente os testes concentraram-se em testar as diversas dimensões das imagens (eixo x) empregando 64 passos de tempo, conforme descrito na Figura 12 (a). O tempo, disposto no eixo y, é medido em segundos e demonstra que imagens pequenas aplicadas ao modelo em 64 passos gastam menos que 1 segundo no processo. Uma figura maior, com dimensão 512×512, obteve um tempo maior que 2 segundos para concluir a cifragem. Na Figura 12 (b) empregando-se o dobro de passos de tempo (isto é, $T = 128$), há um aumento significativo nas imagens maiores e pequenas modificações no tempo total das imagens menores. Antes alcançando pouco mais de 2 segundos, a figura de dimensão 512×512 aplicada a este teste leva mais que 4 segundos para ser cifrada. Ainda assim, este é um tempo pequeno para uma imagem com esta dimensão, haja vista a configuração da máquina.

Ambos os experimentos ilustrados anteriormente, concentraram-se na etapa de ci-

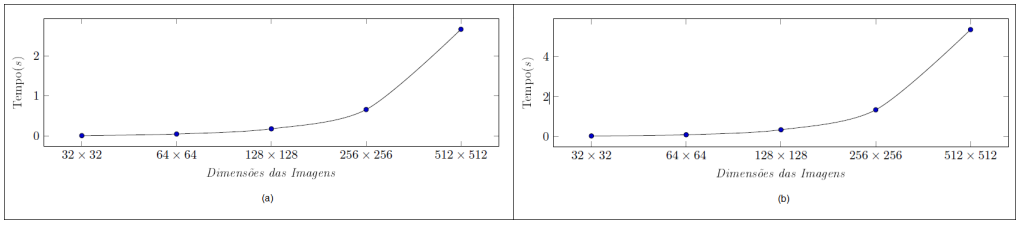


Figura 12. (a) Imagens cifradas em $t = 64$ passos de tempo. (b) Imagens cifradas em $t = 128$ passos de tempo.

fragem. A fase da decifragem também deve ser avaliada uma vez que é constantemente executada nas soluções que envolvam o uso da criptografia. O tempo necessário para decifrar as imagens cifradas no Figura 12 (a) é mostrado na Figura 13 (a), utilizando a mesma quantidade de passos de tempo. Note que a o comportamento do gráfico de tempo de decifragem é similar ao comportamento do tempo da cifragem, porém, a decifragem gasta uma quantidade menor e tempo para ser executada. O mesmo pode ser observado quando são aplicados $T = 128$ passos na decifragem, e a Figura13 (b) mostra esse comportamento.

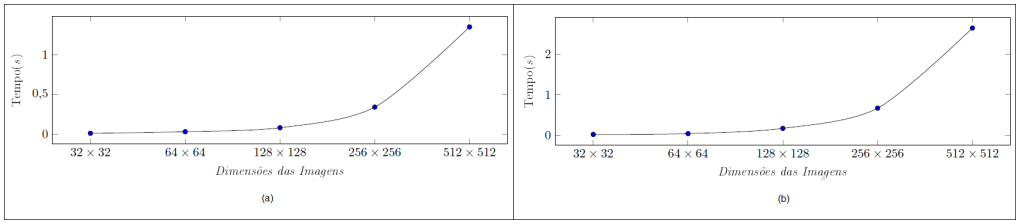


Figura 13. (a) Imagens decifradas em $T = 64$ passos de tempo. (b) Imagens decifradas em $T = 128$ passos de tempo.

5.4 Análise da qualidade da cifragem para imagens coloridas

O histograma é uma ferramenta gráfica capaz de focalizar a tonalidade das cores distribuídas nas imagens. O eixo x e y representam, respectivamente, a tonalidade das cores existentes e a quantidade de pixels empregados a cada cor. Utilizando-se das cores primárias em escalas de 0 a 255, o histograma consegue captar o grau de distribuição de cada um dos canais. A análise do histograma demonstra o esquema de pixels após os processos de pré-embaralhamento e cifragem dos dados. Os histogramas gerados nesta seção foram elaborados através da aplicação do modelo criptográfico em imagens de 64×64, empregando uma regra de raio 2 ($r = 2$) sensitiva à esquerda em 128 passos de tempo ($T = 128$). Os gráficos ilustrados foram elaborados através da ferramenta de manipulação de imagens GIMP [8].

A Figura 14 demonstra a distribuição de cores nos pixels encontrados na primeira imagem analisada. As variações de cada canal estão representadas pela legenda juntamente à coloração representativa da figura. Nota-se que os padrões de cor contidos em cada gráfico formam a imagem original.

Em seguida, os gráficos representados pela Figura 15, demonstram o efeito propor-

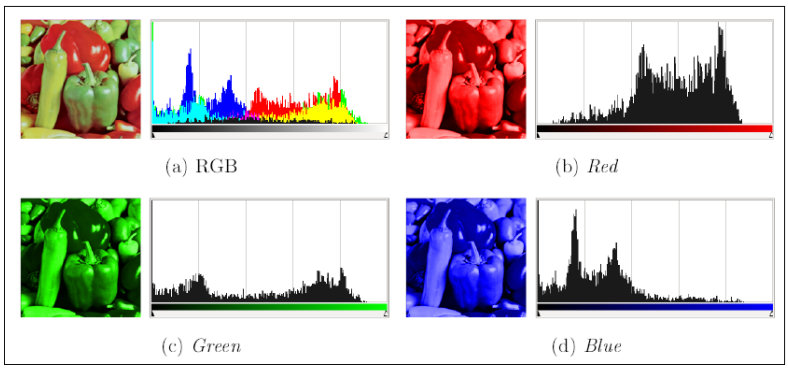


Figura 14. Histogramas de cor e análise de propagação da perturbação na imagem original.

cionado aos pixels, após a execução do pré-embaralhamento descrito anteriormente. Ob-

serve que os níveis de cores apresentados são distorcidos, formando novas figuras. O pré-embaralhamento situa-se no começo do processo de cifragem. Todavia, nota-se uma textura que revela a imagem original o que não é suficiente como forma de cifragem, por isso, o método baseado em ACs unidimensionais caóticos e com borda fixa é aplicado ao pré-embaralhemento para prover a segurança necessária ao mesmo.

O histograma final contido na Figura 16, demonstra o texto cifrado gerado através da

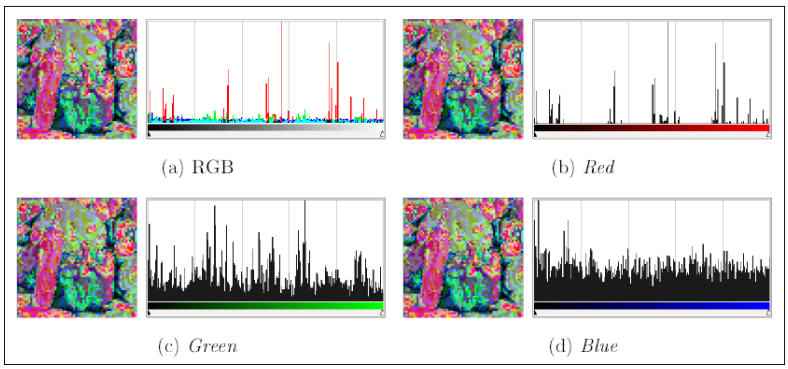


Figura 15. Histogramas de cor e análise de propagação da perturbação em $T = 1$.

abordagem de pré-embaralhamento e posteriormente cifragem de dados através dos algoritmos que foram aqui propostos. Neste caso, é possível observar que a distorção proporcionada pelo embaralhamento anterior, somada à fase de encriptação, acarreta em uma cifra de imagem com padrões divergentes comparados à imagem original. Através desta comparação, é possível expressar com um maior detalhamento a diferença entre o texto claro e o texto cifrado. Essa cifragem apresenta uma boa qualidade, visto que nenhum padrão está associado à imagem original e os histogramas de cores apresentam faixas de valores bem distribuídos em todas os canais de cores do padrão RGB. Embora o canal B esteja com faixas de valores mais esparsas, a cifragem resultante é de boa qualidade, pois alterou muito a dinâmica e a distribuição da imagem original.

5.5 Análise do tempo de processamento de imagens coloridas

O gráfico disposto na Figura 17 (a), demonstra que a maioria das imagens menores ou iguais a 256×256 levaram menos que 10 segundos no processo de cifragem, empregando 64 passos de tempo. Porém, a figura que possui dimensão 512×512 gastou mais que 30 segundos para finalizar o processo. Posteriormente, os novos testes visam computar o tempo total de processamento realizado na decifragem. Inicialmente a Figura 17 (b), demonstra a decifragem dos textos cifrados gerados na Figura 17 (a). Os resultados não diferem muito do tempo total de cifragem. No caso das imagens coloridas, o tempo final de processamento entre as operações de cifragem e decifragem, foram praticamente idênticas. Contudo, o modelo

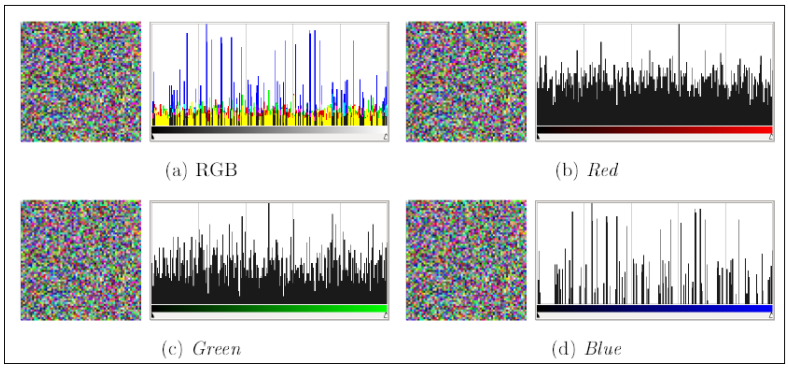


Figura 16. Histogramas de cor e análise de propagação da perturbação em $T = 64$.

apresentou um tempo elevado quando lida com imagens maiores numa abordagem sequencial. Consequentemente, isto reforça que o modelo deve ser implementado numa arquitetura paralela, tornando a criptografia mais rápida aos arquivos maiores.

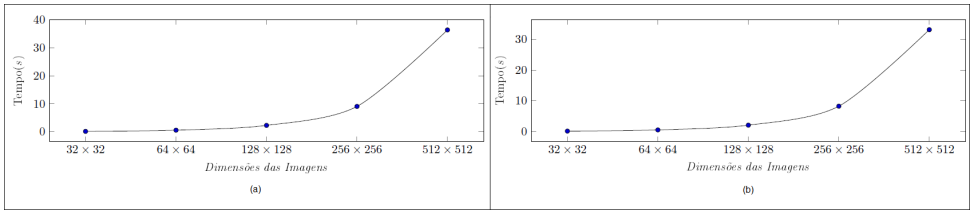


Figura 17. (a) Imagens cifradas em $T = 64$ passos de tempo. (b) Imagens decifradas em $T = 64$ passos de tempo.

6 Discussão dos Resultados

Os experimentos realizados com o modelo de criptografia BCCA tiveram como objetivo avaliar a qualidade da cifragem. O estudo de caso apresentado neste trabalho para a avaliação da cifragem foi a cifragem de imagens digitais coloridas. O modelo BCCA é capaz de cifrar qualquer sequência de bits, incluindo-se textos ou imagens preto e branco, portanto um tratamento de pré-processamento de imagens coloridas foi necessário. Esse pré-processamento foi o responsável por misturar os bits dos canais de cores e também por evitar zonas de texturas Prasad [19] nas imagens cifradas. Além disso, foram realizados experimentos para analisar a qualidade de imagens preto e branco que são representadas por matrizes onde cada pixel é representado por um bit. As maiores conclusões e observações dos experimentos são: (a) o tempo de processamento do método BCCA, se implementado numa arquitetura paralela, em relação ao modelo criptográfico AES é superior, tanto na cifragem

quanto na decifragem; (b) para uma boa cifragem do modelo BCCA - média de 0.5 0s a partir de análise da perturbação - devem ser utilizadas chaves sensíveis com alta entropia $s > 0.7$; (c) a propagação da perturbação pode ser vista ao longo de todo o reticulado, diferentemente do modelo de Gutowitz [9]; (d) o modelo de criptografia BCCA reduz o crescimento acelerado de bits observados nos modelos de Gutowitz [9] e Oliveira [17]; (e) toda sequência de bits como entrada pode ser cifrada utilizando-se o modelo BCCA, diferentemente do modelo Oliveira [18]; (f) o método aqui proposto pode permitir alto nível de paralelismo quando empregado num hardware paralelo, diferentemente de Oliveira [15]; (g) a qualidade de cifragem não é alterada na implementação paralela ou sequencial; (h) o pré-embaralhamento das matrizes de bits para cada canal de cor reduz a criação de zonas de texturas Prasad [19] e portanto foi empregado; (i) o histograma de cores indica que o método BCCA é melhor quando o método é aplicado por mais passos de tempo T; (j) testes de processamento de tempo sequencial foram empregados para mostrarem que à medida que a se aumenta o tamanho da imagem a ser cifrada, aumenta-se também o tempo de processamento tanto no implementação sequencial quanto na paralela, já que ambas são sensíveis ao tamanho da imagem. No entanto, a abordagem paralela do BCCA teria um tempo muito inferior, se comparado a abordagem sequencial, justificando a importância do emprego dos ACs na criptografia.

7 Definição formal do modelo de criptografia BCCA

A criptografia do modelo BCCA pode ser definida por uma 6-tupla $(\mathcal{P}, \mathcal{K}, T, \mathcal{C}, \varepsilon, \mathcal{D})$ que satisfaz as seguintes condições:

1. $\mathcal{P}$ é um conjunto de reticulados iniciais de tamanho $m \times (n + p)$ bits, também denominados blocos originais;

2. T representa a quantidade de passos necessários à cifragem, com $T = \max(m, n, p)$;

3. $\mathcal{K}$ é o conjunto das chaves possíveis $k \in \mathcal{K}$ é composta por uma palavra de 64 bits, sendo o tamanho da regra (chave). A partir do núcleo de uma regra principal sensível a um dos extremos (esquerda ou direita) e com entropia recomendada superior a 0.7;

4. $\mathcal{C}$ é um conjunto de textos cifrados;

5. Para cada $k \in \mathcal{K}$ tem uma regra de criptografia $e_k \in \varepsilon$, dada pelo cálculo de pré-imagem do modelo de AC 1D com borda fixa, e uma correspondente regra de descriptografia $e_k \in \mathcal{D}$, dada pela evolução para frente do modelo de AC 1D com borda fixa. Cada $e_k : \mathcal{P} \to \mathcal{C}$ e $d_k : \mathcal{C} \to \mathcal{P}$ são funções tais que $d_k(e_k(\mathcal{P})) = \mathcal{P}$.

8 Conclusões e trabalhos futuros

Este trabalho apresenta e investiga um novo modelo de criptografia paralelo baseado no cálculo de pré-imagens de autômatos celulares unidimensionais com chaves caóticas e borda fixa, denominado BCCA. Nosso modelo tem como caso de estudo as imagens coloridas digitais, o qual é muito relevante ao contexto de criptografia Prasad [19] (i) o estudo de cifragem de imagens digitais tem ganhado espaço devido ao advento dos equipamentos de captura dessas imagens e constante transmissão desses dados; (ii) criptografia de imagens é um problema canônico à criptografia, pois sua complexidade é superior que a cifragem de um texto e a metodologia se aproxima da cifragem de vídeos, mas com complexidade inferior.

Utilizando-se do cálculo de pré-imagens dos ACs unidimensionais, o modelo de criptografia obteve êxito em cifrar e decifrar imagens digitais. O grau de confusão de bits contidos nas chaves criptográficas de caráter caótico proporciona o efeito de embaralhamento contemplado pelas cifras, seja pelo aspecto visual ou através da análise de propagação da perturbação. O método consegue alcançar um dos principais requisitos relativos à segurança digital atribuído a confidencialidade, onde os textos cifrados originados no processo permitem a recuperação de suas informações somente por entidades autorizadas.

O modelo BCCA emprega como base a criptografia de Gutowitz [9] e Oliveira [17] baseada em ACs. O modelo elaborado diminui a quantidade de bits enviados ao texto cifrado através de uma otimização, representada pela técnica da borda. Além da redução dos bits enviados, o procedimento garante bloquear o crescimento do texto cifrado [9] e [17], aumentando a viabilidade do modelo criptográfico e eliminando problemas, tais como maior gasto de espaço em disco e maior custo no transporte das cifras. Todavia, o modelo ainda contempla uma taxa de bits extras, que deve ser salva para possibilitar a posterior recuperação do texto claro. Esses bits extras são necessários para que todo texto plano dado como entrada possa ser cifrado por qualquer chave (propriedade de sensitividade), diferentemente do algoritmo implementado em [18] que não permite a cifragem de qualquer sequencia de bits de entrada. Além disso, o modelo aqui empregado é altamente paralelizável, diferentemente do modelo de [15], que embora aumente o reticulado apenas quando é necessário, o mecanismo de pilha empregado neste algoritmo, para a recuperação de pré-imagens, reduz o paralelismo do mesmo. Um importante diferencial do modelo em comparação a outros algoritmos conhecidos na literatura, tais como AES e DES Stinson [23] é a capacidade de empregar a computação paralela. Mesmo que alguns métodos tenham sido investigados na tentativa de paralelizar os procedimentos redundantes dos modelos AES e DES Le [10], sabe-se que os mesmos apresentam vários trechos que precisam ser processados de modo sequencial.

Como continuidade deste trabalho, testes de criptoanálise para a verificação se a borda facilita o trabalho de criptoanálise podem ser investigados. Para a melhoria do modelo, caso o texto cifrado seja corrompido, uma análise de perda de dados poderá ser estudada. Uma mudança da atual borda utilizada para uma sequência pseudo-aleatória proposta por Wolfram [26], que partir de um reticulado e um índice gera seus valores aleatoriamente e em seguida

emprega uma execução *forward* em $T = 64$ passos de tempo, é uma outra proposta que pode apresentar ao modelo resultados interessantes. Em resumo, a estrutura da borda é flexível e pode proporcionar diversos experimentos com o objetivo de obter melhores cifragens.

Outra sugestão de alteração no processo de cifragem pode ser a rotação do núcleo das regras de transição (isto é, a geração de novas regras ao processo a partir de modificações do núcleo de uma chave principal). Sabe-se que toda a motivação do emprego do cálculo de pré-imagens baseado em ACs reside na implementação paralela, e que esse procedimento não altera o resultado final da cifragem. Portanto, espera-se aplicar melhorias em relação ao tempo de processamento a partir da implantação na plataforma FPGA (Field Programmable Gate Array) Anghelescu [1], Rajagopalan [20], Raut [21] através da exploração da computação paralela, justificando toda a motivação do emprego de ACs na criptografia.

Contribuição dos autores:

- Eduardo Cassiano da Silva: implementação do modelo proposto BCCA, estudo e validação do modelo, escrita do artigo, implementação e validação dos experimentos, adaptação do banco de imagens para os experimentos.

- Jaqueline Aparecida Jorge Papini Soares: Auxílio e tomada de decisões em relação aos testes e experimentos conduzidos, auxílio na implementação do modelo proposto BCCA, escrita do artigo, orientação do aluno Eduardo Cassiano da Silva.

- Danielli Araújo Lima: Proposição do modelo BCCA, auxílio e tomada de decisões em relação aos testes e experimentos conduzidos, auxílio na implementação do modelo proposto BCCA, escrita do artigo, orientação do aluno Eduardo Cassiano da Silva.

Referências

[1] P. Anghelescu, S. Ionita, and E. Sofron. Fpga implementation of hybrid additive programmable cellular automata encryption algorithm. In *Hybrid Intelligent Systems, 2008. HIS'08. Eighth International Conference on*, pages 96–101. IEEE, 2008.

[2] H. Barros de Macedo, G. M. Barbosa de Oliveira, and C. H. Costa Ribeiro. Dynamic behaviour of network cellular automata with non-chaotic standard rules. In *Complex Systems (WCCS), 2014 Second World Conference on*, pages 451–456. IEEE, 2014.

[3] A. P. Castro and D. A. Lima. Autômatos celulares aplicados a modelagem de dinâmica populacional em situação de risco. *Workshop of Applied Computing for the Management of the Environment and Natural Resources*, 2013.

[4] J. Daemen and V. Rijmen. Rijndael/aes. In *Encyclopedia of Cryptography and Security*, pages 520–524. Springer, 2005.

[5] D. Das and A. Ray. A parallel encryption algorithm for block ciphers based on reversible programmable cellular automata. *arXiv preprint arXiv:1006.2822*, 2010.

[6] C. Feliciani and K. Nishinari. An improved cellular automata model to simulate the behavior of high density crowd and validation by experimental data. *Physica A: Statistical Mechanics and its Applications*, 451:135–148, 2016.

[7] G. B. Ferreira, P. A. Vargas, and G. M. Oliveira. An improved cellular automata-based model for robot path-planning. In *Advances in Autonomous Robotics Systems*, pages 25–36. Springer, 2014.

[8] G. Gimp. Image manipulation program. *User Manual, Edge-Detect Filters, Sobel, The GIMP Documentation Team*, 8(2):8–7, 2008.

[9] H. Gutowitz. Cryptography with dynamical systems. *Kluwer Academic Press*, 1995.

[10] D. Le, J. Chang, X. Gou, A. Zhang, and C. Lu. Parallel aes algorithm for fast data encryption on gpu. In *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, volume 6, pages V6–1. IEEE, 2010.

[11] H. A. Lima and D. A. Lima. Autômatos celulares estocásticos bidimensionais aplicados a simulação de propagação de incêndios em florestas homogêneas. *Workshop of Applied Computing for the Management of the Environment and Natural Resources*, 2014.

[12] M. J. L. Lima. Criptografia baseada no calculo generico de pre-imagens de autômatos celulares. Master's thesis, Universidade Presbiteriana Mackenzie, 2005.

[13] H. B. Macêdo, G. Oliveira, and C. H. Ribeiro. A comparative study between the dynamic behaviours of standard cellular automata and network cellular automata applied to cryptography. *International Journal of Intelligent Systems*, 31(2):189–207, 2016.

[14] J. Machicao, A. G. Marco, and O. M. Bruno. Chaotic encryption method based on life-like cellular automata. *Expert Systems with Applications*, 39(16):12626–12635, 2012.

[15] G. M. Oliveira, L. G. Martins, L. S. Alt, and G. B. Ferreira. Exhaustive evaluation of radius 2 toggle rules for a variable-length cryptographic cellular automata-based model. In *Cellular Automata*, pages 275–286. Springer, 2010.

[16] G. M. B. Oliveira. Autômatos celulares: aspectos dinâmicos e computacionais. *III Jornada de Mini-cursos em Inteligência Artificial (MCIA) - Sociedade Brasileira de Computação*, 8:297 – 345, 2003.

[17] G. M. B. Oliveira, A. Coelho, and L. Monteiro. Cellular automata cryptographic model based on bi-directional toggle rules. *Int. J. of Modern Physics C*, 2004.

[18] G. M. B. Oliveira, M. Lima, H. Macedo, and A. Branquinho. A cryptographic modelo based on the pre-image computation of cellular automata. *Theory and Applications of Cellular Automata*, pages 139 – 155, 2008.

[19] V. C. Prasad and S. Maheswari. Robust watermarking of aes encrypted images for drm systems. In *Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on*, pages 189–193. IEEE, 2013.

[20] S. Rajagopalan, H. N. Upadhyay, J. B. B. Rayappan, and R. Amirtharajan. Dual cellular automata on fpga: An image encryptors chip. *Res. J. Inform. Technol*, 6:223–236, 2014.

[21] L. Raut and D. H. Hoe. Stream cipher design using cellular automata implemented on fpgas. In *System Theory (SSST), 2013 45th Southeastern Symposium on*, pages 146–149. IEEE, 2013.

[22] S. Sen, C. Shaw, D. R. Chowdhuri, N. Ganguly, and P. P. Chaudhuri. Cellular automata based cryptosystem (cac). In *Information and Communications Security*, pages 303–314. Springer, 2002.

[23] D. R. Stinson. *Cryptography: theory and practice*. CRC press, 2005.

[24] S. Vasantha, N. Shivakumar, and D. S. Rao. A new encryption and decryption algorithm for block cipher using cellular automata rules. *International Journal*, 130, 2015.

[25] S. Wolfram. *Cellular Automata.* Los Alamos Science., 1986.

[26] S. Wolfram. *A New Kind of Science*. Wolfram Media - (1st edition): 1197 - 2006-09-19T07:35:05.000+0200, January 2002.

[27] A. Wuensche and M. Lesser. *The global dynamics of cellular automata: An atlas of basin of attraction fields of one-dimensional cellular automata*. Number 1. Andrew Wuensche, 1992.

[28] L. Yu, Y. Li, and X. Xia. Image encryption algorithm based on self-adaptive symmetrical-coupled toggle cellular automata. In *Image and Signal Processing, 2008. CISP'08. Congress on*, volume 3, pages 32–36. IEEE, 2008.

[29] M. Zeghid, M. Machhout, L. Khriji, A. Baganne, and R. Tourki. A modified aes based algorithm for image encryption. *International Journal of Computer Science and Engineering*, 1(1):70–75, 2007.

[30] Y. Zhang, J. Qiao, B. Wu, W. Jiang, X. Xu, and G. Hu. Simulation of oil spill using ann and ca models. In *Geoinformatics, 2015 23rd International Conference on*, pages 1–5. IEEE, 2015.

Método Computacional para o Diagnóstico Precoce da Granulomatose de Wegener

Computational Method for early Diagnosis Wegener's Granulomatosis

José do Nascimento Linhares [1 2]
Lúcio Flávio A. Campos [1 3]
Ewaldo Eder Carvalho Santana [1 4]
Jardiel Nunes Almeida [1 5]
Flávia Larisse da Silva Fernandes [1 6]

Data de submissão: 29/12/2015, *Data de aceite: 25/04/2016*

Resumo: Neste trabalho é apresentado um sistema de reconhecimento de padrões proteômicos com o objetivo de auxiliar o diagnóstico precoce da Granulomatose de Wegener (GW), uma vasculite idiopática rara de difícil detecção e alta taxa de mortalidade para indivíduos não tratados. O método consiste em extrair características de sinais proteômicos e classificá-las como sendo de indivíduos portadores ou não portadores de GW. Para tanto, utiliza-se Análise de Componentes Independentes para extrair características dos sinais, Algoritmo de Máxima Relevância e Mínima Redundância para reduzir o número de características e custos computacionais e Máquina de Vetores de Suporte para classificar. A qualidade do método foi avaliada utilizando uma base de dados com 335 sinais proteômicos, composta por 75 casos ativos, 101 casos negativos e 159 em remissão. O melhor resultado obtido foi para um vetor de vinte características cuja acurácia, especificidade e sensibilidade foram, respectivamente, de: $98,24\%$, $99,73\%$ e $99,50\%$.

***P*alavras-chave:** diagnóstico, granulomatose de Wegener, método computacional, padrões proteômicos

[1]Universidade Estadual do Maranhão (UEMA), Centro de Ciências Tecnológicas, Programa de Pós-Graduação em Engenharia de Computação e Sistemas - São Luís - Maranhão - Brasil
[2]{linhares.jose@yahoo.com.br}
[3]{lucioflavio@gmail.com}
[4]{ewaldoeder@gmail.com}
[5]{jardieliguaiba@gmail.com}
[6]{larisse.nandes@gmail.com}

Abstract: This paper presents a recognition system of proteomic patterns in order to assist in the early diagnosis of Wegener's Granulomatosis (WG), a rare idiopathic vasculitis difficult to detect and of high mortality rate for untreated individuals. The method consists of extracting features of proteomic signs and classifying them as being of bearers individuals or non-carriers of GW. For this purpose, we use Independent Components Analysis to extract characteristics of these signals, Algorithm of Maximum Relevance and Minimum Redundancy to reduce the number of features and computational costs and Support Vector Machine to qualify them. The performance of the method was evaluated using a database of 335 proteomic signals, comprising 75 active cases, 101 negative cases and 159 in remission. The best result was obtained for a vector with twenty characteristics whose accuracy, sensitivity and specificity were, respectively: 98.24%, 99.73% and 99.50%.

***K*eywords:** diagnosis, Wegener's granulomatosis, computational method, proteomic patterns

1 Introdução

A Granulomatose de Wegener (GW) é uma vasculite granulomatosa autoimune multissistêmica rara de difícil detecção, que atinge 3 em cada 100.000 pessoas no mundo (1, 2, 3). Esta doença afeta os vasos sanguíneos de pequeno e médio calibre e vênulas do sistema respiratório superior, pulmões e rins, causando inflamação e consequente necrose dos tecidos desses órgãos. Em alguns casos, pode atingir também o coração, o sistema nervoso, olhos, pele, trato gastrointestinal e musculoesquelético (4, 2). A GW é uma patologia que quando não diagnosticada e tratada precocemente, pode levar o paciente a óbito em apenas um ano.

Atualmente a GW é diagnosticada através de sintomas, exames clínicos, radiológicos e sorológicos que seguem critérios propostos pelo *American College of Rheumatology* (5). Se dois dos seguintes achados: inflamação oral ou nasal, nódulos ou opacidades na radiografia de tórax, hematúria microscópica, inflamação granulomatosa na biópsia da parede de vasos e a presença do anticorpo Anti Citoplasma de Neutrófilos (ANCA-c) positivo forem encontrados, tem-se até 90% de especificidade. Porém, outras doenças da classe das vasculites sistêmicas também apresentam o ANCA-c positivo (6). Vale ressaltar, que os sintomas iniciais da GW são praticamente inespecíficos, o que não permite sua diferenciação em estágios iniciais.

O tratamento é feito com uso de citotóxicos e imunossupressores para combater as reações imunológicas do organismo. O sucesso da terapia está diretamente relacionado com a detecção precoce da enfermidade, pois isto influencia na dosagem dos medicamentos. Se o tratamento for iniciado de forma tardia, doses maiores de medicamentos são aplicadas o que pode potencializar seus efeitos colaterais, trazendo complicações cardíacas, infertilidade, obesidade, osteoporose, hipertensão arterial, diabetes e infecções oportunistas (7). Verifica-se

assim, a necessidade do desenvolvimento de métodos de diagnósticos para a GW que sejam precisos e que permitam a detecção precoce da mesma.

Recentemente a comunidade científica vem aplicando técnicas de CAD (*Computer Aided Diagnosis*) em várias doenças (8, 9, 10, 11). Araújo (8), por exemplo, utiliza a Análise de Componentes Independentes (ICA) para extrair características de sinais proteômicos com o objetivo de diagnosticar o câncer de ovário . Áurea (9) propõe um método de diagnóstico precoce da Diabetes utilizando ICA e Máquina de Vetor de Suporte (SVM). Yu (10) aplica sinais proteômicos e bioinformática para detecção do câncer de colo retal. Mantini (11) usa ICA e padrões proteômicos para identificação de biomarcadores e sua possível associação com doenças.

Neste trabalho, a partir do estudo da espectrometria de massa, especificamente de sinais proteômicos, combinado com métodos computacionais, propõe-se uma metodologia de detecção precoce da GW. O método proposto consiste basicamente em extrair caraterísticas de sinais proteômicos para classificá-los como sendo de indivíduos portadores ou não portadores de GW.

2 Metodologia Proposta

O método proposto é constituído de três submétodos que consistem em: extrair características de sinais proteômicos utilizando Análise de Componentes Independentes (ICA), reduzir a quantidade de características com a técnica de Máxima Relevância e Mínima Redundância (mRMR), afim de diminuir os custos computacionais e classificar com a Máquina de Vetores de Suporte (SVM). A figura 1 mostra um diagrama do método proposto. A seguir descreveremos cada um desses métodos.

2.1 Espectrometria de Massa e Sinais Proteômicos

De acordo com Araújo (12), a ciência tem procurado e desenvolvido formas de diagnosticar doenças precocemente. Nesse sentido o estudo de sinais proteômicos, que é o conjunto de proteínas expressas a partir de um determinado genoma, tem se mostrado promissor, pois o proteoma está em constante mudança devido as respostas que podem ser obtidas aos estímulos externos e internos . Assim, a presença de uma doença pode mudar de forma significativa as características das proteínas e consequentemente do proteoma, indicando qual a patologia que acomete o paciente ou possíveis biomarcadores que possam indicar a presença da enfermidade (13).

Atualmente um dos métodos mais utilizados para obtenção de sinais proteômicos é a espectrometria de massa, que é uma técnica analítica física que permite detectar e identificar moléculas por meio de sua razão massa/carga (m/z). Para a aplicação dessa técnica,

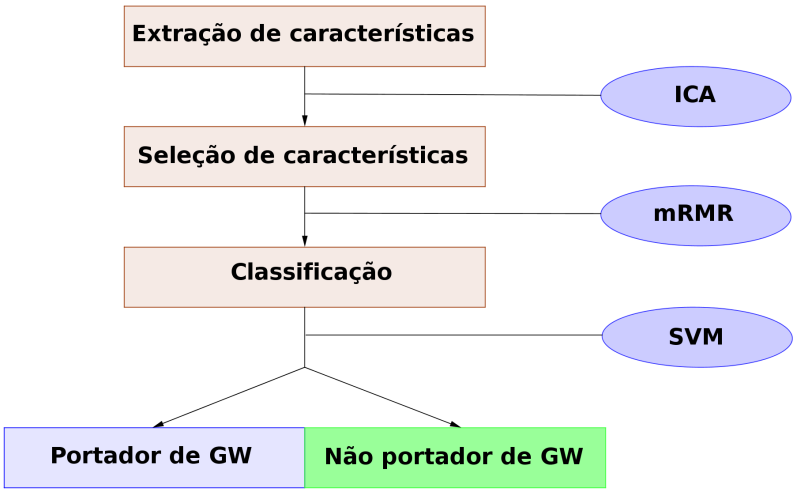


Figura 1. Diagrama da metodologia proposta.

utiliza-se um espectrômetro de massa que é composto basicamente por uma fonte de íons, um analisador de massas, um detector de íons e uma unidade de aquisição de dados.

Neste trabalho utilizamos uma base de dados com sinais proteômicos obtidos a partir de um espectrômetro de massa que utiliza a técnica de ionização *Surface-enhanced laser desorption/ionization* (SELD) e um analisador de massas do tipo *Time of Flight* (TOF) (14). Em SELD, a ionização é feita depositando-se a mistura de proteínas, que se deseja analisar, sobre uma superfície com afinidade química, em seguida, essa superfície é lavada restando apenas as moléculas que se ligaram a ela. Após a lavagem, uma matriz é posta sobre a superfície e deixada cristalizar. Logo após, o analito é excitado por laser para formar os íons em fase gasosa.

No analisador TOF, os íons são acelerados por um potencial elétrico em um tubo de vácuo e detectados de acordo com seu tempo de voo (15), que é proporcional a m/z. O resultado ao final de todo o processo é um espectro de massas. O espectro obtido é um gráfico que mostra a intensidade relativa de cada íon que aparece como picos com m/z definidos. A figura 2 mostra um espectro de massa obtido com a técnica SELD-TOF.

2.2 Extração de Características pela Análise de Componentes Independentes

A análise de componentes independentes (ICA-*Independent Component Analysis*) é um modelo estatístico usado em processamento de sinais para recuperar fontes estatisticamente independentes ou extrair características de um sinal (16). No modelo ICA linear é considerado que um dado vetor aleatório $\mathbf{X}$ de sinais observados, por exemplo, o sinal pro-

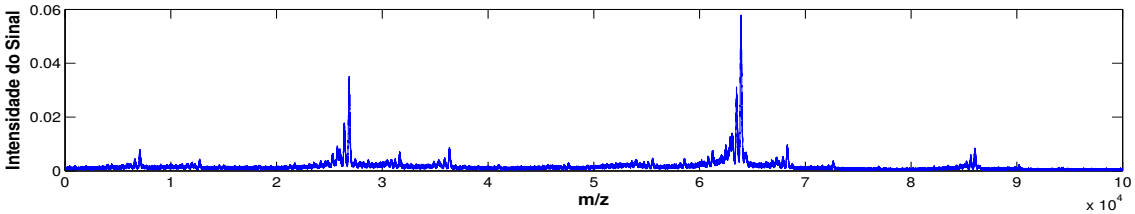


Figura 2. Espectro de massa obtido de um espectrômetro de massas.

teômico, é gerado a partir da atuação de um operador linear $\mathbf{A}$ sobre um vetor $\mathbf{S}$, cujas componentes são mútua e estatisticamente independentes e não gaussianas. Matematicamente pode-se escrever

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

Sendo: $\mathbf{X} = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{pmatrix}$, $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$ e $\mathbf{S} = \begin{pmatrix} s_{11} \\ s_{12} \\ \vdots \\ s_{1n} \end{pmatrix}$.

A matriz $\mathbf{A}$ é vista como uma matriz de mistura e a equação 1 (modelo ICA) mostra como os sinais observados $\mathbf{X}$ são gerados a partir da mistura das componentes independentes de $\mathbf{S}$.

O problema principal em ICA é encontrar $\mathbf{A}$ e $\mathbf{S}$ conhecendo apenas o vetor $\mathbf{X}$ e dependendo da aplicação que se queira fazer, a matriz de interesse poderá ser $\mathbf{A}$ ou $\mathbf{S}$. Na extração de características de sinais proteômicos, por exemplo, a matriz utilizada é $\mathbf{A}$, pois suas colunas representam as características de cada um dos sinais.

Na prática é impossível resolver com exatidão a equação 1 e obter a matriz de características $\mathbf{A}$, porém estimativas podem ser obtidas utilizando a informação mútua ou explorando a propriedade de não gaussianidade das componentes de $\mathbf{S}$. Essa segunda abordagem, tem como alicerce o teorema do limite central, que diz que a soma de variáveis aleatórias estatisticamente independentes e identicamente distribuídas tende a uma distribuição gaussiana (17). Assim, $\mathbf{X}$ tem distribuição de probabilidade mais próxima de uma distribuição gaussiana, uma vez que é gerada pela soma dos elementos de $\mathbf{S}$ ponderados pelos elementos de $\mathbf{A}$.

Para estimar as componentes independentes e a matriz de características $\mathbf{A}$ utiliza-se a equação 1. Nessa equação basta multiplicar os dois lados por $\mathbf{W} = \mathbf{A}^{-1}$ para encontrar $\mathbf{Y} = \mathbf{WX}$, sendo $\mathbf{Y}$ a estimativa de $\mathbf{S}$. Como $\mathbf{X}$ é mais gaussiano que $\mathbf{S}$, uma componente independente é estimada quando se encontra um $\mathbf{W}$ que projeta os elementos de $\mathbf{X}$ em uma

distribuição de probabilidade não gaussiana.

Dentre os algoritmos utilizados para estimar a matriz de características $\mathbf{A}$ e as componentes independentes destaca-se o algoritmo fastICA, por ter rápida convergência, e, comparado com algoritmos baseados em gradiente, é mais simples, pois não necessita de ajuste no passo de adaptação (18). O fastICA usa como medida de não gaussianidade uma versão aproximada da negentropia dada pela equação 2

$$J(y) \approx \sum_{i=1}^{N} k_i [E(G_i(y)) - E(G_i(y_{gaus}))]^2. \tag{2}$$

Sendo os k_i constantes positivas, E é o operador esperança, y_{gaus} variáveis gaussianas com variância unitária e média zero e os G_i são funções não quadráticas. Segundo (19), as funções G_1 e G_2, representadas nas equações 3 e 4, garantem boas aproximações da negentropia e melhoram a convergência do algoritmo fastICA.

$$G_1(y) = \frac{1}{\beta} log(cosh(\beta y)), \text{ com } 1 \leq \beta \leq 2 \tag{3}$$

$$G_2(y) = -exp\left(-\frac{y^2}{2}\right) \ . \tag{4}$$

Os passos de execução do fastICA são:

1. inicializa-se aleatoriamente uma estimativa $\mathbf{W}$ para $\mathbf{A}^{-1}$;
2. ajusta-se $\mathbf{W}$
$$\mathbf{W}_{n+1} \longleftarrow E\{\mathbf{X}G_1(\mathbf{WX}) - G_1'(\mathbf{WX})\}\mathbf{W};$$
G_1' é a derivada de G_1.
3. normaliza-se $\mathbf{W}$
$$\mathbf{W}_{n+1} \longleftarrow \frac{\mathbf{W}_{n+1}}{\|\mathbf{W_{n+1}}\|};$$
4. se não convergir repete-se o passo 2.

Implementações do fastICA nas linguagens R, C++, Python e MATLAB podem ser encontradas em (20).

2.3 Seleção de Características mais Discriminativas

Definir o número de características a serem utilizadas em um sistema de reconhecimento de padrões é de suma importância, pois permite melhorar a performance do classificador, diminuir os custos computacionais e reduzir o tempo na etapa de classificação.

A redução de características consiste na escolha de um subconjunto das características mais informativas produzidas a partir dos sinais originais sem que se perca sua capacidade discriminante (21), isto é, o subconjunto selecionado deve ser capaz de descrever o conjunto como um todo.

Nesse trabalho, foi utilizado o algoritmo de Máxima Relevância e Mínima Redundância (mRMR) para reduzir o conjunto de características. O mRMR seleciona do conjunto **A** as características mais relevantes e menos redundantes. Para tanto, utiliza a medida de máxima relevância, dada pela informação mútua I entre a variável de classe c e cada característica x_i, como mostra equação 5,

$$max\ D(A,c),\ D = \frac{1}{\mid A \mid} \sum_{x_i \in A} I(x_i;c), \tag{5}$$

e minimiza a medida de redundância, uma vez que é possível que entre as características selecionadas via máxima relevância tenham informações redundantes (21, 22) e estas não acrescentam nenhuma informação nova, por isso, podem ser removidas do conjunto de características sem comprometê-lo. A mínima redundância é dada em termos da informação mútua I por 6

$$min\ R(A),\ R = \frac{1}{A^2} \sum_{x_i, x_j \in A} I(x_i;x_j). \tag{6}$$

Em resumo, o mRMR combina as equações 5 e 6 para encontrar a equação 7 que fornece conjuntamente, após um processo de otimização, as características mais relevantes e menos redundantes. Essa equação foi utilizada por Ding e Peng (22) para implementar o algoritmo de máxima relevância e mínima redundância. Tal algoritmo foi testado com varias bases de dados e em todas mostrou-se ser o mais eficiente (22).

$$max\ \Phi(D,R),\ \Phi(D,R) = D - R \tag{7}$$

2.4 Classificação com a Máquina de vetores de suporte

Como etapa final, foi realizada a classificação das amostras utilizando a Máquina de Vetor de Suporte (SVM), que é uma técnica de aprendizado de máquina baseada na teoria do aprendizado estatístico, criado por Vapnick em 1965 para resolver problemas de regressão e classificação (23).

Essa técnica estabelece princípios que permitem induzir um classificador para separar duas ou mais classes de forma que a distância das margens seja máxima. Isso faz com que a SVM seja robusta diante de dados com grandes dimensões, tenha boa capacidade de generalização e suporte ruídos nos dados (24). Aplicações de SVMs podem ser encontradas em categorização de textos, análise de imagens e bioinformática (25).

Para dados linearmente separáveis, um classificador SVM toma como entrada um conjunto de dados e prediz através de uma função de decisão (hiperplano), induzida a partir de um conjunto de treinamento, a que classe cada dado pertence. Em geral o conjunto usado para o treino é um subconjunto das características escolhidas mediante algum critério de seleção como o mRMR. No treino da máquina apenas os dados localizados às margens das classes são utilizados, tais dados são denominados vetores de suporte.

Nas situações em que os elementos do conjunto de dados não sejam linearmente separáveis, a SVM faz o mapeamento desses dados para um espaço de dimensão maior. Nesse espaço, existe uma alta probabilidade que sejam classificados por um hiperplano (26). As funções que realizam a mudança do espaço de representação dos dados do conjunto a ser classificado são chamadas de kerneis.

A tabela 1 mostra as funções kerneis mais utilizadas e que apresentam bons resultados em processos de classificação. Nesse trabalho foi utilizado o kernel definido pela função de base radial (kernel gaussiano).

Tabela 1. Kernel.

Tipo de função	Forma matemática
Função de base radial	$k(x_i, x_j) = e^{-\gamma \lvert x_i - x_j \rvert^2}$
Função polinomial	$k(x_i, x_j) = (1 + x_i \cdot x_j)^n$
Função sigmoidal	$k(x_i, x_j) = tanh(ax_i \cdot x_j + b)$

2.5 Métricas de Desempenho

A avaliação da qualidade de testes diagnósticos é feita, em geral, calculando-se as medidas de acurácia, sensibilidade e especificidade. A acurácia é a taxa de acertos do teste. A sensibilidade é a capacidade que o teste diagnóstico apresenta de detectar os indivíduos verdadeiramente positivos, isto é, de diagnosticar corretamente os doentes. A especificidade informa a eficácia do método em diagnosticar corretamente os indivíduos sadios.

Essas medidas dependem da quantidade de indivíduos classificados correta e incorretamente. Os resultados da classificação podem ser divididos em: verdadeiro positivo, falso positivo, verdadeiro negativo ou falso negativo. Um resultado é definido como verdadeiro

positivo ou verdadeiro negativo se a classificação é feita de forma correta e falso positivo ou falso negativo se ela apresenta resultado incorreto.

As equações para calcular a sensibilidade, a especificidade e a acurácia são, respectivamente (27):

$$Acurácia = \frac{V_P + V_N}{V_P + V_N + F_P + F_N} \tag{8}$$

$$Sensibilidade = \frac{V_P}{V_P + F_N} \tag{9}$$

$$Especificidade = \frac{V_N}{V_N + F_P} \tag{10}$$

Sendo: V_P o número de verdadeiros positivos, V_N o número de verdadeiros negativos, F_P o número de falsos positivos e F_N o número de falsos negativos identificados pelo método.

3 Resultados e Discussão

3.1 Base de dados

Para testar a eficiência desse método, utilizou-se uma base de dados com 335 sinais proteômicos, que pode ser encontrada em (28). Esses sinais foram obtidos por meio da técnica SELDI-TOF e estão divididos em 75 casos com diagnóstico positivo (grupo ativo), 101 casos com diagnóstico negativo (grupo controle) e 159 casos com a doença em fase de remissão. Cada vetor dessa base possui dimensão de 380000. Nesse trabalho, foram utilizados o grupo ativo e o grupo controle.

A figura 3 mostra um sinal proteômico dessa base de dados. O eixo horizontal corresponde aos valores de razão *massa/carga* e o eixo vertical equivale a intensidade do sinal. Cada pico observado dá uma ideia da abundancia das moléculas que compõem esse espectro de massa.

3.2 Extração de características

Como primeira etapa, antecedendo a extração de características, foi realizado um preprocessamento sobre o conjunto de sinais da base de dados, com o objetivo de reduzir os ruídos verificados que certamente degradariam o desempenho do classificador SVM. Nesse

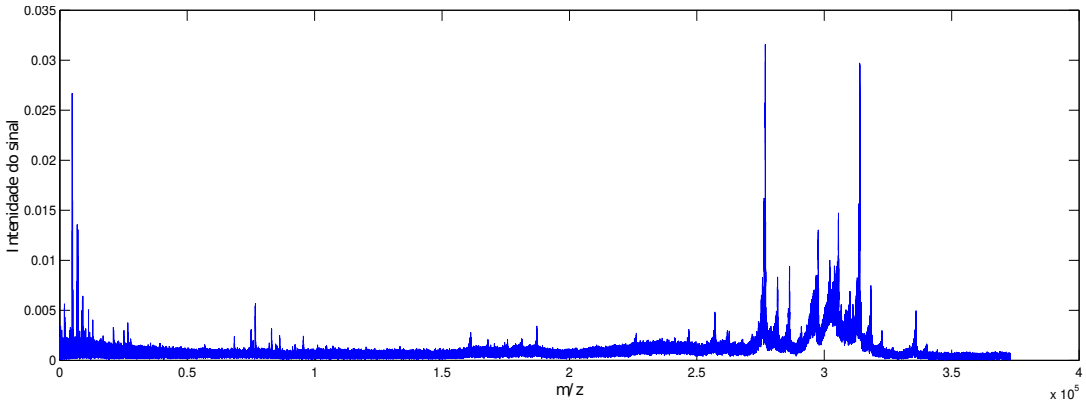


Figura 3. Espectro de massa de um sinal proteômico retirado da base de dados de forma aleatória.

primeiro processo, foi selecionado de cada amostra os pontos no intervalo $[250000; 350000]$, pois verificou-se que a maior parte da informação de todos espectros encontravam-se nesse intervalo.

A figura 4 ilustra os resultados obtidos para dois sinais proteômicos com diagnósticos positivo e negativo, respectivamente, antes e depois desse processo.

O processo de extração de características consistiu em unir os vetores de casos ativos com os vetores de casos negativos, já reduzidos, para gerar a matriz $\mathbf{X}$ de ordem 176×100001 a ser utilizada como entrada no modelo ICA. Cada linha dessa matriz corresponde a um caso e cada coluna a um nível de intensidade do sinal proteômico. Na etapa seguinte, foi utilizado o algoritmo FastICA para extrair as características dos sinais da matriz $\mathbf{X}$. Assim, obteve-se a matriz de características $\mathbf{A}$ de ordem 176×176. As linhas dessa matriz correspondem aos vetores de características dos sinais e permitem identificar cada uma das amostras entre presença ou ausência de Granulomatose de Wegener.

3.3 Redução de dimensionalidade

A redução da dimensionalidade da matriz de características $\mathbf{A}$ foi feita utilizado o algoritmo de Máxima Relevância e Mínima Redundância. Como resultado foi obtido a matriz $\mathbf{A_R}$ com as características organizadas da mais relevante para a menos redundante. Isso significa que as entradas dessa matriz possuem os dados distribuídos em ordem decrescente de representatividade, o que possibilita definir o número de características a serem utilizadas no classificador SVM para obter o seu melhor desempenho.

Para determinar quantas características permitiam um melhor desempenho do classifi-

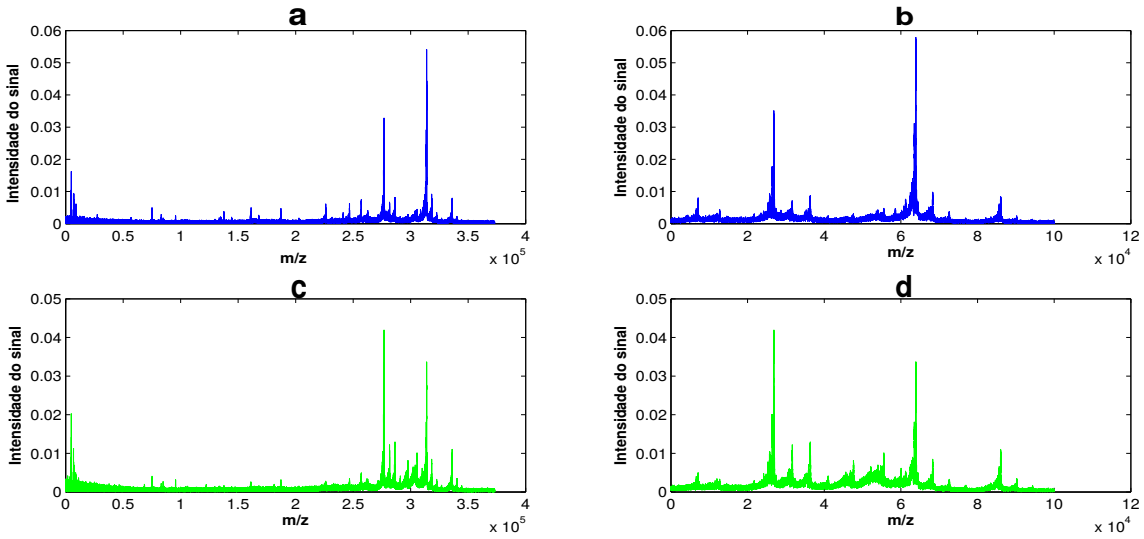


Figura 4. Espectros de massa. A figura (a) corresponde a uma amostra da base de dados com diagnóstico negativo de dimensão 380000 e a figura (b) mostra essa mesma amosta já reduzida para o intervalo $[250000; 350000]$. De forma semelhante, afigura (c) apresenta uma amostra com diagnóstico positivo e a figura (d) equivale a essa amostra com dimensão menor.

cador para cada amostra, foram realizados testes incrementando de cinco em cinco o número de características até um total 175 e cada vetor gerado foi testado com o classificador SVM.

3.4 Classificação das amostras e avaliação do método

Como etapa final, as linhas da matriz $\mathbf{A_R}$, que correspondem aos casos de pacientes portadores e não portadores da GW, foram classificadas por meio da máquina de vetores de suporte, utilizando o kernel dado pela função de base radial representada na tabela 1, com $\gamma = 0,5$.

Por último, foi avaliada a eficácia do método proposto calculando a acurácia, a sensibilidade e a especificidade do classificador com a técnica de validação cruzada *10-fold-cross validation*, que consistiu em dividir a base de dados em dez partes, usar nove para treino e uma para teste. Esse processo foi repetido permutando circularmente as divisões até que todas fossem usadas.

A tabela 2 mostra os melhores resultados obtidos no processo de classificação pela SVM. Estes foram alcançados com vetores de 5, 10, 15 e 20 características. Da observação desses dados é possível ver que o melhor desempenho do classificador e, consequentemente,

do método proposto foi obtido para um vetor com 20 características (linha 4 da tabela 2). Para esse vetor, obteve-se $98,24\%$ de acurácia, $99,73\%$ de sensibilidade e $99,50\%$ de especificidade, com desvios padrão respectivamente de $0,174$, $0,035$ e $0,073$. Isso significa que dos 176 indivíduos portadores e não portadores de GW, 173 foram diagnosticados corretamente (soma dos verdadeiros positivos V_P com os verdadeiros negativos V_N) e 3 de forma incorreta (soma dos falsos positivos F_P com os falsos negativos F_N). Apenas um indivíduo foi diagnosticado como normal (falso negativo) sendo portador de GW.

Tabela 2. Desempenho da SVM para 5, 10, 15 e 20 características. A acurácia, a sensibilidade e a especificidade são apresentadas com seus respectivos desvios padrões.

Carac	VP	FP	VN	FN	Acurácia	Especificidade	Sensibilidade
5	73	3	98	2	(97,22±1,94)%	(97,93±3,24)%	(96,33±2,43)%
10	73	2	99	2	(97,75±2,07)%	(98,28±2,66)%	(98,70±2,30)%
15	73	2	99	2	(97,75 ±1,97)%	(94,85±3,86)%	(99,10±1,62)%
20	**74**	**2**	**99**	**1**	(**98,24** $\pm 1,74$)%	(**99,73** $\pm 0,35$)%	(**99,50** $\pm 0,73$)%

Para implementação da metodologia proposta foi utilizada a linguagem de programação *MatLab*, utilizando os pacotes fastICA e mRMR, disponíveis em (20) e (29), respectivamente, e o pacote SVM, foi adquirido de (8).

4 Considerações Finais

Neste trabalho foi apresentado um método computacional que utiliza Análise de Componentes Independentes, técnica de seleção de atributos Máxima Relevância e Mínima Redundância e Máquina de Vetores de Suporte para diagnosticar precocemente a Granulomatose de Wegener, uma doença rara com complicações multissistêmica que quando não diagnosticada e tratada rapidamente pode levar o paciente a morte. Esse método foi usado para classificar 176 sinais proteômicos de pacientes e os resultados corroboram estudos anteriores quanto à eficiência da técnica ICA para extrair características de sinais proteômicos, a mRMR permite selecionar as melhores características que identificam os portadores de GW, além de reduzir custos computacionais e a SVM implementada com um kernel gaussiano tem um bom desempenho num cenário de classificação não linear.

Para um vetor com apenas vinte características o método proposto obteve $98,24\%$ de acurácia, $99,73\%$ de sensibilidade e $99,50\%$ de especificidade. Das 176 amostras apenas 3 foram classificadas incorretamente, sendo duas falso positivo e uma falso negativo.

Apesar dos bons resultados, para um aumento da confiabilidade do método apresentado novos testes devem ser realizados em diferentes bases de dados.

Diante dos resultados apresentados, espera-se que em um futuro bem próximo o método desenvolvido neste trabalho possa ajudar profissionais da saúde no diagnóstico da Granulomatose de Wegener. Isso possibilitará um aumento da sobrevida do paciente com diagnóstico positivo, uma vez que a completa remissão dessa doença está relacionada com a precocidade do tratamento.

Contribuição dos autores:

Os autores contribuíram de forma equivalente na construção do presente artigo.

Referências

[1] REZENDE, C. E. B. et al. Granulomatose de wegener: relato de caso. *Revista Brasileira de Otorrinolaringologia*, v. 69, n. 2, p. 261–265, 2003. ISSN 1809-4570. Disponível em: <http://www.scielo.br/pdf/rboto/v69n2/15634.pdf>. Acesso em: 2 mar. 2014.

[2] FIGUEIREDO, S. et al. Granulomatose de wegener: Envolvimento otológico, nasal, laringotraqueal e pulmonar. *Revista Portuguesa de Pneumologia*, v. 15, n. 5, p. 929–935, 2009. ISSN 0873-2159. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2173511509701630>. Acesso em: 27 abr. 2014.

[3] SANTOS, S. K. J. dos et al. Granulomatose de wegener: importância do diagnóstico precoce. relato de caso. *Revista Brasileira Clinica Medica*, v. 7, p. 427–433, 2009. ISSN 1679-1010. Disponível em: <http://www.sbcm.org.br/revista/completas.php>. Acesso em: 02 set. 2014.

[4] GOMIDES, A. P. M. et al. Perda auditiva neurossensorial em pacientes com granulomatose de wegener: Relato de três casos e revisão de literatura. *Revista Brasileira de Reumatologia*, v. 46, n. 3, p. 234–236, 2006. ISSN 1809-4570. Disponível em: <http://www.scielo.br/pdf/rbr/v46n3/31356.pdf>. Acesso em: 2 mar. 2014.

[5] RHEUMATOLOGY, A. C. of. *Granulomatosis with Polyangiitis (Wegener's)*. 2014. Disponível em: <http://www.rheumatology.org/I-Am-A/Patient-Caregiver/Diseases-Conditions/Granulomatosis-with-Polyangitis-Wegners>. Acesso em: 2 mar. 2014.

[6] RADU, A. S.; LEVI, M. Anticorpos contra o citoplasma de neutrófilos. *Jornal Brasileiro de Pneumologia*, v. 1, n. 31, p. 16–20, 2009. ISSN 1806-3756. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1806-37132005000700006>. Acesso em: 21 abr. 2014.

[7] STONE, J. H. et al. A serum proteomic approach to gauging the state of remission in wegeners granulomatosis. *American College of Rheumatology*, v. 52, n. 3, p. 902–910, 2005. ISSN 2175-2745. Disponível em: <http://seer.ufrgs.br/index.php/rita/article/view/rita_v14_n2_p43-67/3543>. Acesso em: 21 jun. 2014.

[8] ARAUJO, W. B. D.; CAMPOS, L. F. A.; ALINE, S. F. Método de detecção de câncer de ovário utilizando padrões proteômicos, análise de componentes independentes e máquina de vetores de suporte. In: XIV WORKSHOP DE INFORMÁTICA MÉDICA, 14. *Anais do congresso da sociedade brasileira de computação*. Brasília: CSBC, 2014. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/wim/2014/011.pdf>. Acesso em: 2 dez. 2014.

[9] RIBEIRO, A. C. et al. Diabetes classification using a redundancy reduction preprocessor. *Research on Biomedical Engineering*, v. 31, n. 2, p. 97–106, 2015. ISSN 2446-4740. Disponível em: <http://www.rebejournal.org/files/v31n2/v31n2a02.pdf>. Acesso em: 3 jul. 2015.

[10] YU, J. K.; CHEN, Y. D.; ZHENG, S. An integrated approach to the detection of colorectal cancer utilizing proteomics and bioinformatics. *World journal of gastroenterology: WJG*, Baishideng Publishing Group Inc, v. 10, n. 21, p. 3127–3131, 2004. ISSN 2219-2840.

[11] MANTINI, D. et al. Independent component analysis for the extraction of reliable protein signal profiles from maldi-tof mass spectra. *Bioinformatics*, Oxford Univ Press, v. 24, n. 1, p. 63–70, 2008.

[12] ARAUJO, W. B. D. *Método de detecção de câncer de ovário utilizando análise de componentes independentes, algoritmo de máxima relevâcia e mínima redundância e máquina de vetores de suporte*. Dissertação (Mestrado em Engenharia de Computação e Sistemas) — Universidade Estadual do Maranhão, São Luís, 2014.

[13] GALDOS-RIVEROS, A. C. et al. Proteômica: novas fronteiras na pesquisa clínica. *Enciclopédia Biosfera*, v. 6, n. 11, p. 1–24, 2010.

[14] AFONSO, C. et al. Activated surfaces for laser desorption mass spectrometry: application for peptide and protein analysis. *Current pharmaceutical design*, Bentham Science Publishers, v. 11, n. 20, p. 2559–2576, 2005.

[15] WILSON, K.; WALKER, J. *Principles and techniques of biochemistry and molecular biology*. [S.l.]: Cambridge university press, 2010.

[16] DENNER, R. R. G. *Compressão de Sinais de Eletrocardiograma Utilizando Análise de Componentes Independentes*. Dissertação (Programa de Pós-Graduação em Engenharia de Eletricidade) — Universidade Federal do Maranhão, São Luís, 2006.

[17] PAPOULIS, A. (Ed.). *Probability, Random Variables and Stochastic Processes*. New York, USA: McGraw-Hill, 1991.

[18] LEITE, V. C. M. N. *Separação Cega de Sinais: análise comparativa de algoritmos*. Dissertação (Programa de Pós-Graduação em Engenharia Elétrica) — Universidade Federal de Itajubá, Itajubá, 2004.

[19] HYVARINEN, A.; KARHUNEN, J.; OJA, E. (Ed.). *Independent component analysis*. New York: John Wiley e Sons, 2001.

[20] AAPO. *Independent Component Analysis (ICA) and Blind Source Separation (BSS)*. Disponível em: <http://research.ics.aalto.fi/ica/fastica/>. Acesso em: 2 mar. 2014.

[21] CATARINO, F. M. I. F. *Segmentação da íris em imagens com ruído*. Dissertação (Dissertação de Mestrado) — Universidade da Beira Interior, Covilhã, 2009.

[22] DING, C.; PENG, H. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, Imperial College Press, v. 3, n. 2, p. 185–205, 2005. ISSN 1757-6334. Disponível em: <http://penglab.janelia.org/papersall/docpdf/2004_JBCB_feasel-04-06-15.pdf>.

[23] GUNN, S. *Support Vector Machines for Classification and Regression*. 1998. Disponível em: <http://users.ecs.soton.ac.uk/srg/publications/pdf/SVM.pdf>. Acesso em: 2 set. 2014.

[24] RODRIGUES, T. A. O. et al. Predição de função de proteínas através da extração de características físico-químicas. *Revista de Informática Teórica e Aplicada*, v. 22, n. 1, p. 29–51, 2015. ISSN 2175-2745. Disponível em: <http://seer.ufrgs.br/index.php/rita/article/view/RITA-VOL22-NR1-29/33912>. Acesso em: 2 jul. 2015.

[25] LORENA, A. C.; CARVAHO, A. C. P. L. F. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. ISSN 2175-2745. Disponível em: <http://seer.ufrgs.br/index.php/rita/article/view/rita_v14_n2_p43-67/3543>. Acesso em: 21 abr. 2014.

[26] HAYKIN, S. (Ed.). *Redes neurais: princípios e prática*. Porto Alegre: Bookman, 2007.

[27] NEVES, S. C. F. *Classificação de câncer de ovário através de padrão proteômico e análise de componentes independentes*. Dissertação (Programa de Pós-Graduação em Engenharia de Eletricidade) — Universidade Federal do Maranhão, São Luís, 2012.

[28] PROGRAM, C. P. *Biomarker Profiling, Discovery and Identification*. 2015. Disponível em: <http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>. Acesso em: 2 mar. 2015.

[29] MATWORKS. *minimum-redundancy maximum-relevance feature selection*. 2015. Disponível em: <http://www.mathworks.com/matlabcentral/fileexchange/14916-minimum-redundancy-maximum-relevance-feature-selection>. Acesso em: 6 mar. 2015.

ZipfTool: Uma ferramenta bibliométrica para auxílio na pesquisa teórica

ZipfTool: A bibliometric tool for supporting in theoretical research

Diego Nunes Molinos [1]
Daniel Gomes Mesquita [2]
Debora Nayar Hoff [2 3]

Data de submissão: 14/10/2015, *Data de aceite: 13/05/2016*

Resumo: Devido ao volume de trabalhos publicados nos veículos de divulgação científica, ferramentas de análise de dados textuais tornam-se importantes para diversas áreas de conhecimento. Utilitários dessa natureza oferecem ao usuário funcionalidades tanto em âmbito quantitativo quanto qualitativo. Do ponto de vista quantitativo, possibilitam identificar a frequência de ocorrência de palavras no texto e diferenciar verbos, substantivos e artigos definidos. Já as análises qualitativas tratam do levantamento de palavras de maior conteúdo semântico. Este trabalho tem por finalidade apresentar o desenvolvimento de uma ferramenta de análise de dados que não somente possui primitivas de análise quantitativas mas também qualitativas. Quantitativamente a ferramenta fornece a frequência dos principais termos do texto, enquanto qualitativamente ela identifica as palavras de maior teor semântico. Utilizando de técnicas advindas da bibliometria, a ferramenta apresentada, chamada de ZipfTool implementa tanto a 1^a quanto a 2^a Leis de Zipf. Este trabalho também apresenta um estudo de caso na área de arquitetura de computadores e mostra uma redução do universo de artigos a serem analisados de 46785 para 1508, permitindo observar a importância da utilização da ferramenta ZipfTool principalmente no auxilio para observação de conceitos, termos e palavras.

***P*alavras-chave:** lei de Zipf, análise textual, bibliometria, conteúdo semântico, frequência de ocorrência, análise qualitativa, análise quantitativa, computação reconfigurável

[1]Universidade: Universidade Federal de Uberlândia, UFU - Uberlândia, Minas Gerais, Brasil. {diego.molinos@ufu.br}
[2]Universidade Federal do Pampa, UNIPAMPA - Santana do Livramento, Rio Grande do Sul, Brasil. {mesquita@unipampa.edu.br}
[3]{deborahoff@unipampa.edu.br}

Abstract: Due to the high number of scientific publications, textual data analysis tools are important for many knowledge fields. Such tool's features can be both in quantitative and qualitative levels. On one hand quantitative tools allows to identify the frequency of of words in the text and differentiate verbs, nouns and definite articles. On the other hand, qualitative tools analyzes identifies words of greater semantic content, identifies descriptors and keywords in the text. This study aims to present the development of a data analysis tool with both quantitative and qualitative approaches. Quantitatively the tool provides the frequency of the main terms of the text, while qualitatively it identifies the words of greater semantic content. Using techniques arising from bibliometrics, the presented tool, so called ZipfTool both implements the 1^a and 2^a Laws of Zipf. This article also presents a case study in the field of computer architecture that allows us to see the importance of using the ZipfTool, mainly with respect to a more accurate observation of concepts, terms, words and definitions.

***K*eywords:** Zipf's law, textual analysis, bibliometric, semantic content, occurrence frequencies, quantitative analysis, qualitative analysis, reconfigurable computing

1 Introdução

A busca pela construção do conhecimento remete para a importância de estabelecer-se uma discussão na direção do campo teórico. Dentro do campo teórico, conceitos e termos são apresentados como construções lógicas, os quais são estabelecidas de acordo com um sistema de referência [20]. O uso indiscriminado de termos e conceitos dentro das ciências, de modo geral, conduz ao empobrecimento do campo de estudo e dos próprios conceitos.

Pode-se, segundo [1] observar que a noção de $conceito$ tem-se confundido muito com a noção de $significado$, resultando em uma analogia errônea com o conceito de objeto, tendo a ideia de algo pronto, passível de ser apenas decorado e repetido. Sócrates mostrou que a definição conceitual se inicia com o raciocínio indutivo, expressando a essência ou a natureza de algo [22]. Segundo [8], o conceito é constituído de elementos que se articulam numa unidade estruturada. Dentre esses elementos há enunciados verdadeiros sobre o objeto que se deseja descrever. Esses enunciados são expressos através de signos que possam traduzir e fixar os enunciados que definem um objeto. Comumente esses signos são palavras. Disso abstrai-se a necessidade de identificação do peso semântico das palavras em um texto.

Ainda conforme [8], se no quotidiano a imprecisão na descrição de objetos pode não trazer grandes consequências, quando se trata de linguagens mais especializadas, essas consequências podem ser desagradáveis. Um exemplo de "linguagens mais especializadas"é a forma de expressão no meio científico, onde conceitos são utilizados para compreensão da

realidade e para replicação de experimentos. Neste espaço não há, ou pelo menos não deveria haver, margem para subjetividade.

Para [31], um dos maiores obstáculos ao desenvolvimento do conhecimento humano advém justamente da imprecisão dos termos utilizados na constituição dos saberes. Ainda segundo a autora, esta dificuldade gera confusões e inadequações de graves consequências.

Por outro lado, o processo de construção do conhecimento normalmente se dá dentro de um paradigma científico. Segundo [18] durante um período de tempo, a evolução científica acontece dentro de parâmetros aceitos pela comunidade de pesquisadores de um tema ou área. Entretanto o autor salienta que muitas das grandes revoluções científicas ocorrem em dissonância com os paradigmas vigentes, causando rupturas que podem levar à construção de novos paradigmas.

Em todo caso, seja para enquadrar-se no paradigma atual, seja para refutá-lo, o cientista precisa ler e compreender os trabalhos relacionados com o seu. Entretanto, considerando as facilidades das publicações digitais e da internet, a tarefa de seleção e leitura do que é relevante pode ser árdua[4]. Ainda que se faça uso de filtros disponibilizados pelas editoras de publicações eletrônicas ou ferramentas de busca, o número de artigos relacionados com determinadas palavras-chave pode ser elevado, dificultando o discernimento do pesquisador.

Uma forma de otimizar o tempo do pesquisador seria a utilização de técnicas que o permitissem analisar e catalogar, automaticamente, artigos baseados em termos específicos nos textos. Tais termos podem ser descritores ou palavras-chave relevantes para seu campo de pesquisa. Desta forma, artigos com conteúdo semântico menos relevante para o tema da pesquisa podem ser descartados sem a necessidade de uma leitura completa desses artigos, poupando tempo para textos mais significativos.

Técnicas advindas da bibliometria permitem analisar a frequência de ocorrência de palavras dentro de um texto, lançando mão de métodos matemáticos e estatísticos para investigar e quantificar os processos de comunicação e escrita. Neste contexto, destaca-se a lei de [33], que além calcular a frequência de ocorrência das palavras dentro do texto, permite a identificação de palavras-chaves e descritores, bem como as palavras de maior conteúdo semântico dentro de um determinado texto.

Este trabalho tem como objetivo apresentar uma ferramenta auxiliar para pesquisa científica, baseada na Lei de Zipf, bem como discutir um estudo de caso relacionado com pesquisa em arquitetura de computadores.

Para uma melhor compreensão deste trabalho, a seção 2 apresenta os fundamentos matemáticos da Lei de Zipf, além de mencionar o ponto de transição de Goffman. Já a seção

[4]Em 13 de abril de 2015 a IEEE Xplore Digital Library comemorava a chegada ao número de dois milhões de artigos publicados em HTML.

3 discute a importância da utilização de ferramentas de análise textual, bem como apresenta alguns aplicativos que implementam a Lei de Zipf nesse contexto. A ferramenta ZipfTool, proposta neste trabalho, é descrita na seção 4 . Um estudo de caso é discutido na seção 5.2, no qual a ferramenta foi utilizada para auxiliar na criação de um arcabouço conceitual para pesquisas na área da computação reconfiguráveis. Finalmente, a seção 6 discute os resultados obtidos e fornece um panorama dos trabalhos atuais e futuros relacionados com a ferramenta ZipfTool.

2 Lei de Zipf e o ponto de transição de Goffman

A bibliometria é a área do conhecimento que utiliza métodos matemáticos e estatísticos para investigar e quantificar os processos de comunicação e escrita [13]. Convergente a este conceito, [16] indica que a bibliometria compreende o exame dos aspectos quantitativos dos processos de produção, disseminação e uso da informação registrada, contendo medidas e modelos matemáticos que auxiliam os exercícios de prospecção e tomada de decisão. Desses modelos quantitativos, a Lei de [33] é uma técnica que identifica a frequência de ocorrência de palavras dentro de um texto longo. Como exercício, Zipf analisou a obra Ulisses, de James Joyce. Percebeu então uma correlação entre a frequência em que um termo aparecia e sua posição na lista de palavras ordenadas segundo sua frequência de ocorrência. Isso levou Zipf a concluir que havia uma regularidade na seleção e no uso das palavras. Também observou que a posição de um termo, multiplicada por sua frequência iguala-se a uma constante (≈ 26.500). Essa lei pode ser expressa como:

> "O produto da ordem de série (R) de uma palavra pela sua frequência de ocorrência (F) é aproximadamente constante (C).

A "ordem de série"é a representação temática da organização das palavras em ordem, de acordo com a quantidade de vezes que elas aparecem no texto. Isso significa que a palavra de maior número de ocorrências tem ordem 1, e que vem logo em seguida tem ordem 2, e assim por diante. Matematicamente, a Lei de Zipf pode ser descrita como na Equação 1.

$$R \times F = C \tag{1}$$

Entretanto foi observado que essa lei não se aplica para palavras de baixa frequência. O próprio Zipf propôs uma segunda Equação para tratar dessa anomalia. Essa Equação foi revisada e modificada por [2], que deu a forma da Equação 2.

$$\frac{I_1}{I_n} = \frac{n \times (n+1)}{2} \tag{2}$$

Na Equação 2, I_1 representa a quantidade de palavras que tem frequência 1, I_n representa a quantidade de palavras que tem frequência n, e 2 é a constante válida para a língua inglesa.

A constante válida empregada na equação 2 trata-se de um valor utilizado para análise em textos escritos na língua inglesa, existem trabalhos, [6] e [19] que fazem menção a utilização da Lei de Zipf em outros idiomas, porém, fazem uso de abordagens empíricas para adaptar este valor de constante.

De acordo com [19], a Lei de Zipf, em sua forma natural é valida para língua portuguesa, pois, nessa forma primitiva a lei trata do nível de ocorrência de palavras no texto e a indexação das mesmas. A modificação proposta por [2] a princípio não é aplicável para trabalhos escritos na língua portuguesa, porém, ao se alterar o denominador 2 da equação, pelo denominador 1.5, identificado por [19] para textos em língua portuguesa, é possível adequar a modificação proposta por [2] para esta língua.

As Equações 1 e 2 descrevem o comportamento das palavras situadas nas extremidades da lista de distribuição em um dado texto. Portanto pode-se inferir que há uma região com palavras cujas frequências de ocorrência são similares. Nessa região crítica há uma transição de comportamento de palavras de alta frequência para palavras de baixa frequência. [12] levantou a hipótese de que as palavras de maior conteúdo semântico (descritores, palavras-chave ou termos de indexação) de um determinado texto estariam nessa região.

Conforme proposto por [2], palavras que possuem baixa frequência tem seu número de ocorrência tendendo a 1. Então, substituindo I_n por 1 na Equação 2, obtém-se a Equação 3:

$$\frac{I_1}{1} = \frac{n \times (n+1)}{2} \tag{3}$$

Que pode ser rearranjada como:

$$n^2 + n - 2 \times I_1 = 0 \tag{4}$$

Resolvendo-se a Equação 4 através da popular fórmula de Bhaskara [5], levando-se em consideração apenas a raiz positiva, tem-se:

$$n = \frac{-1 + \sqrt{1 - 8 \times I_1}}{2} \tag{5}$$

[5] Ainda que não haja evidência que o brilhante matemático indiano do século XII tenha desenvolvido a resolução de equações de 2^o, essa é a nomenclatura ensinada nas escolas do Brasil.

Então, o n calculado na Equação 5 é denominado ponto de transição (T) de Goffman, que determina graficamente a localização da transição das palavras de alta frequência para as de baixa. Segundo sua hipóteses, existe uma região em torno desse ponto onde há maior probabilidade de encontrar-se as palavras com maior conteúdo semântico. A Figura 1 ilustra a área de transição em torno do ponto T de Goffman, bem como as regiões onde se encontram as palavras de alta frequência de ocorrência e as de baixa.

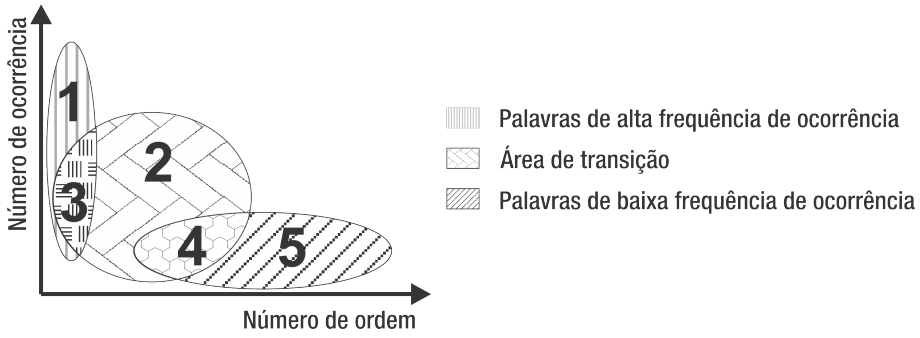


Figura 1. Zonas de ocorrência de palavras classificadas segundo a Lei de Zipf-Booth e interpretadas segundo Goffman

Ainda na Figura 1, a primeira zona de ocorrência, representada na Figura pelo número 1 é composta por palavras de maior número de ocorrências. Essas palavras normalmente são as raízes de sintaxe do idioma em que o texto é escrito (por exemplo, artigos definidos ou indefinidos). Já a segunda zona, descriminado na Figura pelo número 2 se caracteriza por conter uma quantidade maior de representantes de categorias morfológicas e informativas do que a primeira zona, como substantivos, adjetivos e verbos. A terceira zona, representada pelo numeral 3 é conhecida como ponto de transição, que de acordo com Goffman se encontram as palavras de maior teor semântico. Finalmente, a quarta e quinta zona contém instâncias que ocorrem uma única vez.

3 Referencial teórico e trabalhos relacionados

Esta seção ressalta a importância da análise textual e lista algumas ferramentas disponíveis para essa atividade.

3.1 Importância da análise textual

Conforme [24], a leitura e a produção textual são atividades habituais no quotidiano de milhares de pessoas, as quais estão diretamente relacionadas com o desenvolvimento intelectual e social. São consideradas de extrema importância para o aprendizado, no entanto,

são notáveis as dificuldades enfrentadas pelos leitores na tentativa de captar a ideia intrínseca do texto.

De acordo com [5], a análise de conteúdo textual trata-se de métodos qualitativos de extração de dados, sendo esses compreendidos por um conjunto de técnicas que convergem para a busca do entendimento de um determinado texto. Esse processo denomina-se análise semântica.

Ainda de acordo com [5], afirma-se que o método de análise de conteúdo é balizado por duas vertentes: A da linguística tradicional e a da interpretação do sentido das palavras (hermenêutica). A primeira vertente visa guiar o pesquisador a utilizar métodos de análise lógicos e estéticos, onde se tem a busca por aspectos formais típicos do texto. Já a segunda vertente prioriza métodos puramente semânticos, partindo da interpretação epistemológica e ontológica de palavras e frases de um texto.

A complexidade do processo de interpretação de um texto fica evidente, uma vez que exige a compreensão do conhecimento expresso na sua forma escrita, para além da ambiguidade natural de quase todos os idiomas. Ressalta-se aqui a importância das definições precisas dos conceitos no campo teórico, uma vez que favorecem a superação da ambiguidade. Outro aspecto, já citado, que contribui para a complexidade da tarefa diz respeito ao imenso volume de informações disponíveis atualmente. Buscar referencial teórico e trabalhos relacionados para embasar uma pesquisa científica têm se tornado uma tarefa hercúlea.

Considerados esses aspectos, é cada vez mais necessário o uso de ferramentas eficientes para auxiliar os cientistas na construção do conhecimento.

Ferramentas e algoritmos que possuem técnicas de análise de textos são largamente utilizados para extrair, organizar e observar o comportamento do conhecimento através dos textos, oferecendo apoio na identificação de termos relevantes, palavras chaves e descritores inseridos nos textos [17].

Na seção seguinte são listadas alguns aplicativos relacionados com este tema.

3.2 Ferramentas de análise textual

A análise textual, em grande parte dos casos, exige a manipulação de grandes quantidades de dados. Assim, o desenvolvimento de software específico para este fim contribui para a redução do esforço manual, gerando resultados de maneira mais rápida e organizada. Do ponto de vista prático e analítico a análise textual é definida como um método de extração de dados relevantes, utilizando bases de dados não estruturadas, ou semi-estruturadas [10]. Do ponto de vista do software de análise, os mesmo devem atender alguns requisitos funcionais que servem de apoio para a análise. Conforme [17] esses requisitos podem ser definidos como:

1. **Contagem de Termos:** Levantamento do número de termos utilizados no texto, tais como, artigos, preposições, verbos e sujeitos;
2. **Apresentação dos termos:** visualização de todos os termos catalogados;
3. **Apresentação dos termos relevantes:** identificação dos termos relevantes, tais como: descritores, palavras chaves e termos mais utilizados;
4. **Frequência dos termos:** visualização da frequência de ocorrência de cada termo dentro do texto;
5. **Relacionamento entre os termos:** visualização dos relacionamentos entre dois ou mais termos para identificação das ideias gerais do texto;
6. **Visualização gráfica dos termos e relacionamentos entre os termos:** visualização gráfica de todos os termos e relacionamento entre eles;

Através da análise desses requisitos pode-se delinear a diferença entre os conceitos aplicados juntamente com uma melhor compreensão dos termos utilizados, permitindo novas proposições e uma análise diferenciada sobre o texto. Abaixo segue uma breve descrição sobre algumas ferramentas avaliadas frente aos requisitos citados.

3.2.1 *TextAnalyzer*

O *TextAnalyzer* trata-se de uma ferramenta online gratuita que se encontra ativa desde abril de 2009 e foi desenvolvida pela iniciativa *Online-Utility.org* com o objetivo de auxiliar pesquisadores, escritores e alunos na análise de textos. Sob a luz da analise textual, a ferramenta permite várias análises do texto, tais como: identificação da frequência de palavras, identificação do número de palavras, identificação do número de sílabas. Além desses aspectos a ferramenta também consegue identificar frases correlacionando as palavras de maior frequência no texto. A ferramenta possui suporte para vários idiomas e não possui uma interface gráfica amigável, sendo bastante robusta para uma análise textual única, mas torna-se inviável para um grande número de textos.

3.2.2 *WordCounter*

O *WordCounter* é uma ferramenta de análise textual online, desenvolvida por [25] que tem como objetivo principal a mensuração das palavras que possuem maior frequência de ocorrência no texto. Assim como a grande maioria das ferramentas online, a mesma não dispõe de uma interface gráfica para auxilio visual das informações. Uma das aplicações possíveis para a o *WordCounter* é a mensuração de palavras que se encontram repetidas no texto, evitando possíveis repetição de palavras [17]. Assim como outras ferramentas online, oferece suporte para análise unitária e não para um grupo ou conjunto de textos.

3.2.3 *SOBEK*

O *SOBEK* é uma ferramenta desenvolvida com o intuito de servir de apoio para professores e pesquisadores na evolução do processo textual. A ferramenta apresenta aspectos bem atrativos do ponto de vista de usabilidade pois, permite acompanhar todo o processo de escrita de forma clara e inteligível [21]. Conforme [17] a ferramenta apresenta grande potencial para análise textual, principalmente em casos de bases de dados não estruturadas. A ferramenta trabalha de forma unitária, ou seja, analisando um texto por vez e possibilita ao usuário visualização dos principais termos e relacionamentos em forma de grafos, permitindo identificar a ideia original do texto através de conceitos analisados pré-definidos [21].

3.2.4 *TagCrowd*

O *TagCrowd* foi desenvolvido por [29], trata-se de uma ferramenta de análise textual simples, porém robusta, não apresenta muitas funcionalidades para o âmbito da análise de dados, oferecendo o básico da análise quantitativa de palavras [17]. Através de configurações parametrizadas, tais como: o número de palavras que deseja-se obter como resultado, idioma e a frequência de ocorrência, a ferramenta direciona a analise sobre o texto predefinido. Diferentemente de outras ferramentas online, o *TagCrowd* oferece uma interface gráfica mais amigável, ilustrando de forma diferenciada palavras que possuem maior frequência de ocorrência no texto. Assim, como as outras ferramentas online, não possui suporte para analisar vários textos de forma dinâmica, oferecendo apenas análise unitária.

3.2.5 *Uff - Lei de Zipf*

Desenvolvida pelo Instituto de Matemática da Universidade Federal Fluminense a ferramenta *Uff - Lei de Zipf* a ferramenta apresenta robustez no cálculo de frequência de ocorrência das palavras de um determinado texto, juntamente com resultados marginais convenientes para o estudo da Lei de Zipf. A mesma não possui uma interface visual amigável, porém oferece a possibilidade da visualização das informações através de gráficos o que possibilita a interpretação dos resultados de forma diferente. Para que o usuário parametrize os resultados a ferramenta oferece a possibilidade de inserção de filtros na análise. Assim como ferramentas online para esse propósito específico, recai-se sobre a mesma problemática, análise de textos unitária não permitindo análise de textos em lote.

3.2.6 *Iramuteq*

A ferramenta *IRAMUTEQ* trata-se de um software gratuito, desenvolvido por [26] e licenciado pela GNU GPL (v2). O software foi projetado tendo como base o software R (www.r-project.org) e a linguagem Python de programação (www.python.org)[4]. Dentre as principais análises realizadas pelo software, além das análises clássicas, ou seja, análise

lexicográfica e frequência de palavras, a ferramenta permite uma análise mais apurada, como, análise de similitude que utiliza teoria de grafos, possibilitando identificar as coocorrências entre as palavras e da conexão entre as mesmas [4] [26].

A Figura 2 faz uma ilustração das ferramentas apresentadas anteriormente face aos requisitos listados por [17].

Ferramenta	Funcionalidades						
	Online	Contagem de termos	Visualização dos termos	Termos relevantes	Frequência dos termos	Relacionamento dos termos	Visualização gráfica
TextAnalyser	X	X	X		X		
WordCounter	X	X	X		X		
SOBEK	X	X	X	X	X	X	X
TagCrowd	X			X	X		X
UffZipf	X	X	X	X	X	X	X
Iramuteq	X	X	X	X	X	X	X

Figura 2. Ferramentas analisadas face aos requisitos propostos por [17]

Faz-se necessário esclarecer que a Figura 2 foi adaptada dos requisitos funcionais propostos por [17], tendo como modificação mais significativa a coluna de termos relevantes. Para o autor supracitado, termos relevantes são termos que possuem alto nível de ocorrência no texto, discorda-se deste entendimento, pois acredita-se que palavras que possuem maior valor semântico são termos relevantes e nem sempre os mesmos se encontram entre as palavras de maior ocorrência do texto [13], dessa forma nenhuma das ferramentas atendem a esse critério de forma positiva.

4 Ferramenta ZipfTool

A ferramenta ZipfTool foi desenvolvida com o principal objetivo de auxiliar pesquisadores no processo de análise de textos. A ferramenta realiza a análise do texto em sua totalidade, extraindo informações quantitativas e qualitativas do mesmo. Cabe-se salientar que ferramentas de análise de dados textuais não são e não devem ser taxadas como ferramentas de mineração de dados ou como técnica de mineração, pois, conforme [9], o processo de mineração de dados consiste em uma sequência predefinida de tarefas aplicadas sob uma base da dados comum, como exemplo de tarefas pode-se citar: Análise de Regras de Associação, Análise de Padrões Sequenciais, Classificação e Predição, Aglomeração e Análise de Outliers. Diante deste contexto pode-se inferir que ferramenta ZipfTool de análise textual partilha de conceitos como associações de dados, classificação, aglomeração e prognósticos, porém não faz uso de nenhuma técnica de mineração muito menos uso de nenhuma ferramenta projeta para tal tarefa, tais como: Intelligent Miner de [3] ou DBminer de [30].

Projetada no ambiente matemático MATLAB, a ZipfTool realiza a leitura de arquivos com extensão .TXT [6]. Quanto ao formato, trata-se do mesmo utilizado pelas ferramentas *online* já mencionadas. Uma conversão do formato .PDF para .TXT ficou de fora do escopo desta ferramenta, uma vez que existem aplicativos e scripts gratuitos que se ocupam eficientemente dessa tarefa [7].

4.1 Especificação

Conforme [28], especificação de um software está diretamente relacionado a definição de suas funcionalidades, em resumo, o levantamento dos requisitos da ferramenta, não ignorando as restrições e limitações que a mesma possui, pois de fato, esse saldo entre requisitos e restrições é o que caracteriza a especificação de uma ferramenta. Abaixo seguem os principais tópicos que retratam as especificações da ferramenta ZipfTool.

1. Ambiente para execução;
2. Formato de textos suportados;
3. Quantidade e tamanho de textos para análise;e
4. Regras de configuração;

Abaixo é apresentado cada tópico da especificação juntamente com seus requisitos funcionais.

4.1.1 Ambiente para execução: A ferramenta ZipfTool foi desenvolvida utilizando o software MATLAB (Versão 2012a). Dessa forma para que a ferramenta ZipfTool funcione, a plataforma citada deve estar instalada. Não há necessidade de instalação de nenhum pacote adicional na ferramenta para o funcionamento da mesma.

4.1.2 Formato de textos suportados: Devido a plataforma MATLAB não ser exclusivamente projetada para trabalhar com diferentes arquivos do tipo texto, o modelo padrão mais comum e primitivo de texto foi adotado, o **.TXT**, que é reconhecido e manipulado por qualquer sistema operacional e qualquer plataforma de desenvolvimento. Cabe salientar que a plataforma MATLAB possui diversas bibliotecas de software já pré-definidas e configuradas com inúmeras funções que auxiliam o carregamento, análise e armazenamento de arquivos em **.TXT**.

[6] Formato reconhecido por todos os sistemas operacionais e aplicativos.
[7] Como por exemplo o "Free PDF to Text Converter"do fabricante LotApps, que pode ser gratuitamente obtido através do "http://lotapps-free-pdf-to-text-converter.soft32.com"

4.1.3 Idiomas dos textos: A ferramenta ZipfTool foi projetada para analisar trabalhos escritos em língua Inglesa, porém, existe trabalhos baseados em abordagens empíricas que preveem a modificação das equações da Lei de Zipf para adaptar a mesma para trabalhos na língua portuguesa.

4.1.4 Quantidade e tamanho de textos suportados: A ferramenta ZipfTool não possui um limite pré estabelecido com relação ao tamanho dos arquivos para análise, vale lembrar que esse limite está diretamente relacionado as limitações de hardware do computador, como por exemplo, memória *RAM* disponível no momento da execução dos scripts e também o espaço disponível nas mídias de armazenamento em massa (*Hard Disk*) para armazenamento dos resultados. Cabe salientar que o software MATLAB em execução sobre o sistema operacional possui as características de um processo em execução, sendo assim, o mesmo partilha de um tamanho limitado de memória física do hardware para execução de suas tarefas, essa caracterizada como memória virtual, a qual é dedicada ao processo no momento de sua execução. Os scripts do MATLAB carregam todos os arquivos .TXT no momento da execução, sendo o número máximo de arquivos bem como o tamanho dos mesmos diretamente relacionados com estes aspectos apresentados. Diante do contexto, diversos testes foram realizados, os scripts conseguiram carregar arquivos de até 162 páginas ou arquivos fragmentados que somados não ultrapassem 162 páginas.

Diferentemente das ferramentas *online*, as quais possuem a capacidade de análise de um texto por vez, a ferramenta ZipfTool já foi projetada tendo como um dos seus requisitos funcionais a análise em lote para quando se deseja analisar diversos textos. A ferramenta possui a capacidade de organizar os resultados em múltiplos arquivos de saídas.

4.1.5 Regras de configuração: A ferramenta ZipfTool utiliza-se de *scripts* para carregamento, execução e salvamento das tarefas. Por tratar-se de *scripts* os mesmo necessitam de pré-configuração, cujas variáveis de configuração são apresentadas na Tabela 1.

4.2 Desenvolvimento

A ferramenta ZipfTool foi projetada visando preencher algumas lacunas as quais foram observadas em outras ferramentas de análise textual, tais como: análise em lote de arquivos de texto e a identificação de termos de maior conteúdo semântico no texto. As subseções a seguir relatam aspectos do desenvolvimento da ZipfTool.

4.2.1 Plataforma de desenvolvimento

A ferramenta ZipfTool foi inicialmente instituída tendo como base os requisitos funcionais básicos de análise de dados conforme apresentado na seção 3. Como já mencio-

Tabela 1. Descrição das variáveis do arquivo de configuração da ferramenta ZipfTool

Variável de Configuração	Descrição
$MyDir$	Define o caminho ou diretório onde se encontram os arquivos para análise.
$Directory$	Define o caminho ou diretório de saída, usado para armazenar os resultados.
$NumberofChar$	Define um número mínimo de caracteres que as palavras analisadas devem conter para não serem descartadas.
$Exception$	Define um conjunto de palavras, separadas por espaço em branco, que serão descartadas no momento da análise textual.

nado anteriormente, para o desenvolvimento da ferramenta ZipfTool foi utilizado o software MATLAB R2012a, o qual possui um propósito de ser uma plataforma de desenvolvimento interativo, possuindo alto desempenho para cálculos numéricos, assim permitindo desenvolver soluções com otimização de tempo de desenvolvimento em relação a outras plataformas de desenvolvimento dessa natureza. Apesar de não ser uma plataforma própria para implementação de algoritmos de análise textual o mesmo possui diversas bibliotecas e funções já pré-definidas que permitem implementações neste âmbito.

Abaixo são apresentadas algumas características da plataforma MATLAB,

1. **Sintaxe** - O MATLAB utiliza a linguagem M-Code ou simplesmente M. Através de uma forma interativa o usuário interage com a plataforma de duas formas, a primeira dela é utilizando o *Comand Window*, um espécie de prompt de comando, onde o usuário insere os comandos e a ferramenta executa o processamento tendo como base o histórico de comandos já inseridos, podendo também interagir através de scripts, onde o usuário pode definir o cadenciamento dos mesmos e executá-los em lote, [11].

2. **Scripts** - Trata-se de um contentor de comandos, onde cada vez que o script é executado, os comandos são processados de forma *top-down*[8], iniciando na primeira linha e terminando somente na última. A utilização de scripts é bastante conveniente, pois os mesmos permitem alterações e atualizações do código e a possibilidade do mesmo ser executado inúmeras vezes, [11].

3. **Visualização gráfica** - Por se tratar de uma plataforma de desenvolvimento voltada para cálculos numéricos, a plataforma possui diversas formas de ilustração gráfica, partindo do 2D até gráficos tridimensionais.

[8]Método de execução de linhas de código, onde a execução é ordenada linha após linha, iniciando de cima para baixo

Diante do apresentado cabe-se esclarecer que do ponto de vista do ambiente de desenvolvimento, a ferramenta ZipfTool trata-se de um conjunto de scripts MATLAB, ordenados e finitos para realização das tarefas conforme é apresentado neste trabalho.

4.2.2 Scripts

A ferramenta ZipfTool possui 4 scripts desenvolvidos em linguagem M-Code. Um script principal, similar a um programa principal, que é responsável por executar todos os outros scripts como procedimentos. Abaixo segue em detalhes as funcionalidades de cada script.

1. **Script Conf.m:** Trata-se do script responsável pelo carregamento das informações de controle do processo de análise. O script **Conf.m** é o primeiro script a ser executado na cadeia de script, o mesmo é responsável por colher informações do arquivo *conf.txt*, onde tem-se as informações sobre o diretório de entrada (conjunto de textos a serem analisados), diretório de saída (conjunto de resultados da análise) e o número de descarte (palavras que possuem os números de caracteres iguais aos números de descarte são descartadas automaticamente). O script **Conf.m** também faz a leitura do arquivo *exception.txt*, o qual é responsável por armazenar palavras que não serão computadas (descartadas), e que não inferem de forma positiva nos resultados, como exemplo, artigos definidos e indefinidos da gramática.

2. **Script RunMyFiles.m:** Este script tem como principal objetivo fazer uma varredura dentro do diretório estabelecido como universo de dados (definido no arquivo de configuração) e organizar os arquivos para análise, de forma que, quando a análise estiver sendo realizada, todos os textos estejam completamente catalogados.

3. **Script Load.m:** Este script é responsável por fazer o carregamento de todos os caracteres do texto e o armazenamento na memória temporária (cache) do processo de análise, ficando essas disponíveis até o próximo carregamento de dados.

4. **Script Zipf.m:** Este script é responsável por executar (instanciar) todos os scripts explicados anteriormente, aplicar a tratativa básica de padronização (acentuação, maiúsculo e minúsculo) bem como executar de todas as primitivas matemáticas que envolve a 1^a e 2^a Lei de Zipf, calcular de frequência de palavras, sequenciar as palavras conforme nível de ocorrência, calcular o ponto de transição do Goffman bem como identificar as palavras que fazer parte desta érea de transição.

4.2.3 Arquivo de configuração

A ferramenta ZipfTool possui um arquivo de configuração, chamado de *conf.txt*, que é responsável pelas configurações de partida, contendo informações essenciais para o funciona-

mento de toda processo de análise. O script *Conf.m* está programado para ler cada linha deste arquivo como sendo uma configuração, abaixo segue é apresentado a estrutura do arquivo supracitado:

1. A primeira linha do arquivo *Conf.txt* deve ser atualizada com as informações do diretório que contém os arquivos para análise, já no formato **TXT**. O caminho deve ser especificado em sua completude, incluindo partição, pastas e subpastas.

2. A segunda linha do arquivo *Conf.txt* deve ser atualizada com as informações do diretório onde os resultados das análises são armazenados. O caminho deve ser específicado de forma completa, incluindo partição, pastas e subpastas.

3. Terceira linha do arquivos *Conf.txt* deve ser atualizada com o número de caracteres de descarte, esse número é interessante pois, o script Zipf.m armazena o número de caracteres de cada palavras que é catalogada e analisada, assim, as que possuírem o número menor ou igual ao definido são automaticamente descartadas da análise.

4.3 Aplicações da ferramenta

A ferramenta ZipfTool é um mecanismo de análise textual, cabe salientar, que a ferramenta possui implementado as primitivas matemáticas da 1^a e a 2^a Leis de Zipf, permitindo assim, uma análise dos termos de maior valor semântico no texto, além de cálculo da frequência de ocorrência das palavras. A Figura 3, apresenta possíveis aplicações da ferramentas dentro do âmbito da análise textual.

APLICAÇÃO	DESCRIÇÃO	ÁREA DE ATUAÇÃO
Identificação da frequência de ocorrência de palavras no texto	A ferramenta permite identificar a frequência de ocorrência de palavras em um determinado texto. Aplicável em processos, onde se deseja identificar palavras repetidas no texto, palavras de maior ocorrência bem como as de menor ocorrência no texto.	Auxílio nas escritas de trabalhos científicos, artigos e revistas. Análises textuais de uma forma geral.
Identificação de palavras de maior teor semântico	A ferramenta permite identificar palavras de maior teor semântico no texto, as quais são classificadas como descritores, termos relevantes e palavras chave. Aplicável em processos onde necessita-se verificar a natureza do texto, bem como o índice de utilização de termos dentro do texto.	Análises conceituais em artigos, revistas, livros, etc. Auxílio na avaliação terminológica dos termos.
Análise de conjunto de textos	A ferramenta permite a análise sobre um conjunto de textos de forma dinâmica, sem a necessidade de carregamento unitário. Aplicável em processos onde o sucesso está condicionado ao número de trabalhos analisados.	Avaliações conceituais, onde o número de definições comprometem a consistência dos resultados. Avaliação de inúmeros trabalhos em geral onde o aspecto quantitativo é importante.

Figura 3. Possíveis aplicações para ferramenta ZipfTool

5 Resultados e Análise

Esta seção é responsável por apresentar os resultados quantitativos no que tange às funcionalidades da ferramenta ZipfTool em relação a outras ferramentas de análise textual. Além disso, traz um estudo de caso da ferramenta aplicada no auxílio em uma pesquisa na área de computação reconfigurável, onde fica claro a contribuição qualitativa da ferramenta no auxílio de pesquisas em âmbito teórico. Cabe salientar que em nenhum momento foi efetuado testes qualitativos para avaliar a qualidade dos resultados gerados das ferramentas citadas na seção 3, não é objetivo deste trabalho apresentar resultados nesse âmbito.

5.1 Comparação entre a ZipfTool e outras ferramentas

De acordo com o apresentado na seção 3, [17] enfatiza alguns requisitos funcionais os quais as ferramentas de análise de textos devem incorporar.

A Figura 4 ilustra um comparativo entre os requisitos propostos por [17] face as funcionalidades da ferramenta ZipfTool.

Ferramenta	Funcionalidades									
	Online	Contagem de termos	Visualização dos termos	Termos relevantes	Frequência dos termos	Relacionamento dos termos	Visualização gráfica	Análise em Lote	Termos de maior conteúdo semântico	Parametrização da Análise
TextAnalyser	X	X	X		X					
WordCounter	X	X	X		X					
SOBEK	X	X	X	X	X	X	X			
TagCrowd	X			X	X		X			X
UffZipf	X	X	X	X	X	X	X			X
Iramuteq	X	X	X	X	X	X	X			X
ZipfTool		X	X	X	X		X	X	X	X

Figura 4. Análise quantitativa das ferramentas de análise de dados

Pode ser observado que a ferramenta ZipfTool atende a praticamente todos os requisitos apresentados por [17], não abrangendo os requisitos de *Ferramenta OnLine* e *Relacionamento dos termos*. Nota-se que alguns requisitos funcionais os quais não são contemplados no trabalho de [17] são adicionados, pois, entende-se que os mesmo são importantes para análise dos textos, são eles:

1. **Análise em Lote:** Permite a ferramenta analisar diversos arquivos de textos não necessitando o carregamento unitário. Funcionalidade importante para análise de grandes volumes de dados de entrada.
2. **Termos de maior conteúdo semântico:** Discorda-se de [17] no que se refere a termos relevantes, para o mesmo são palavras que possuem maior frequência de ocorrência

dentro do texto. De acordo com [13], palavras de maior conteúdo semânticos são termos relevantes que aparecem no texto e definem a ideia principal do autor, normalmente, palavras chaves, descritores e indexadores, os quais podem auxiliar a definir aspectos morfológicos do texto.

3. **Parametrização das variáveis de análise:** Ilustrando uma análise textual em forma gráfica, gera-se algo similar a Figura 3.2.6, onde pode ser observado que a região onde se localizam as palavras de maior frequência de ocorrência no texto são compostas palavras, em aspectos morfológicos conhecidas como: artigos, conjunções e interjeições; Sendo em uma região mais nobre, denominada de região de transição, encontra-se palavras, em aspectos morfológicos conhecidas como: adjetivos, verbos, preposições, pronomes e sujeitos, sendo esses termos o mais prováveis a serem descritores, palavras chave ou termos de maior conteúdo semântico do texto. Formas de parametrização de análise que possibilitem ignorar um grupo de palavras que possuem um determinado número de letras ou até mesmo um grupo seleto de palavras se torna bem interessante do ponto de vista analítico.

Como pode ser observado através da análise da Figura 4, a ferramenta Zipftool apresenta como principal diferencial a possibilidade de análise de arquivos do tipo lote, ou seja, diversos arquivos sem a necessidade de carregamento um a um, porém também tem como ponto negativo a mesma não estar disponível online e não efetuar relacionamento dos termos de análise.

5.2 Estudo de Caso

O estudo de caso se justifica do ponto de vista científico como uma maneira metódica de descrever o exemplo de um determinado conhecimento [32]. O estudo de caso descrito nessa seção é compreendido por um trabalho de pesquisa realizado na Universidade Federal de Uberlândia, realizado no contexto de um mestrado acadêmico. Seu objetivo foi uma rigorosa e detalhada análise conceitual sobre um tema específico no campo da computação. Abaixo são apresentados maiores detalhes sobre essa dissertação.

5.2.1 Arcabouço Conceitual para Computação Reconfigurável

A pesquisa compreendida pelo título supracitado foi desenvolvida a partir de uma análise feita em artigos publicados na área da computação reconfigurável. Os autores do trabalho constataram que havia uma inconsistência conceitual entre os termos utilizados dentro deste campo de estudo. Após realizarem uma vasta leitura em trabalhos publicados, observaram que alguns termos ora eram tratados como sinônimos e ora tratados como coisas bem divergentes.

Uma discussão conceitual dentro de qualquer campo de estudo, sempre apresenta importância significativa para área, já que "conceitos"são considerados instrumentos fundamentais para compreensão da área. O uso indiscriminado dos conceitos dentro da uma érea científica, de modo geral, conduz ao empobrecimento da mesma [20].

Em um trabalho prévio nosso, [23], para avaliar a qualidade e a forma com que os conceitos estavam sendo empregados, precisou-se desenvolver um processo de análise das principais definições conceituais encontrados em artigos na área da computação reconfigurável. A primeira dificuldade destacada por [23] foi a quantidade de trabalhos publicados que compõem o universo da Computação Reconfigurável, incluindo livros, jornais, revistas e artigos de congressos e outros eventos científicos. A partir de uma análise amostrada, foi observado que nem todos esses trabalhos publicados possuíam definições claras sobre os termos que envolvem a computação reconfigurável, sendo que os artigos lidos pareciam não valorizar a definição conceitual dos termos.

A Tabela 2 mostra a quantidade de trabalhos selecionados através da biblioteca *online* [15]. A localização destes trabalho foi feita usando-se os termos de busca apresentados na Tabela 5, sem o uso de qualquer tipo de filtro adicional.

Tabela 2. Número de publicações selecionados pela ferramenta online

Termo utilizado	Quantidade de Trabalhos selecionados
Reconfigurable Computing	5771
Reconfigurable Hardware	6550
Reconfigurable Architecture	9659
FPGA	24805

A partir dos dados apresentados na Tabela 2, pode-se dizer que o universo a ser analisado no campo sob observação é bastante grande. Mesmo que fossem detectadas redundâncias nos artigos selecionados em cada termos, trata-se de 40 mil trabalhos a serem analisados. Isso posto, fica evidente a necessidade de uma ferramenta que permita a identificação dos trabalhos mais relevantes.

5.2.2 Uso da ZipfTool como ferramenta auxiliar na construção do Arcabouço Conceitual para Computação Reconfigurável

A utilização da ferramenta ZipfTool no estudo de caso apresentado possui duas vertentes, uma qualitativa e outra quantitativa. Sob a luz do aspecto qualitativo, a ferramenta possui a capacidade de identificar os termos de maior conteúdo semântico do texto, como forma de evidenciar os trabalhos convergentes ao objetivo da busca, isso tende a poupar um grande esforço inicial manual de seleção e análise dos trabalhos, o que tende a otimizar o

tempo de pesquisa. No aspecto quantitativo, a ferramenta tende a induzir que sejam selecionados somente trabalhos que possuem alta relevância para com os objetivos de pesquisa, gerando assim, um universo amostral reduzido e de alto valor semântico.

Resultados Qualitativos

Tendo como hipótese inicial de que, muitos dos trabalhos que faziam parte do universo inicial de análise, mesmo após o procedimento de amostragem, não seriam de grande auxilio para a pesquisa. Isso porque, o primeiro levantamento de dados foi realizado utilizando a ferramenta *Online* [15], sendo que a mesma seleciona os trabalhos de sua base de dados correlacionando os termos de busca e os títulos dos trabalhos presentes em sua base de dados, não exercendo verificação de conteúdo sobre os trabalhos.

Os termos de busca instituídos inicialmente apenas faziam referência a termos que são constantemente citados na área de estudo de computação reconfiguravél, podendo ou não, terem relações com os objetivos da pesquisa. Para tanto, foi necessário identificar termos de busca que estimulassem a seleção natural de trabalhos que convergissem com os objetivos da pesquisa. Para tal tarefa, inicialmente foram identificados trabalhos consagrados na área da computação reconfigurável como tendo um ótimo embasamento conceitual e por terem sido construídos por pesquisadores expressivos e reconhecidos no campo, a julgar pelo número citações que os mesmos possuem. Esta estratificação focada elimina certas confusões conceituais e permite um direcionamento mais polido com relação aos potenciais termos de busca, eliminando termos menos impactantes. Sobre os artigos [27], [7] e [14] foi aplicada a ferramenta ZipfTool que selecionou os termos de maior conteúdo semântico dos trabalhos citados. O resultado é o apresentado na Figura 5.

Podem ser observados na Figura 5, número de palavras analisadas, frequência de ocorrência e identificação de palavras de maior teor semântico relacionadas a cada trabalho analisado.

Como resultado desta etapa de análise qualitativa, observa-se que as palavras *reconfigurable* e *architecture* aparecem em duas das três análises como palavras de alto teor semântico juntamente com a palavra *FPGA*. Diante do exposto, estas palavras se tornam os novos termos de busca da pesquisa, pois conforme a análise realizada, as mesmas possuem a capacidade de selecionar trabalhos com conteúdo impactante para a pesquisa.

Cabe-se ressaltar que, esta análise qualitativa com o objetivo de elencar termos de busca que incitem a ferramenta a selecionar trabalhos impactantes para a pesquisa proporcionou aos autores visualizar que, alguns veículos de publicações possuíam um maior número de trabalhos publicados relacionados com os novos termos de busca e outros veículos não. Este processo descrito auxiliou também de forma qualitativa os autores a selecionarem veículos de publicações com maior número de trabalhos impactantes para a pesquisa, tendo como resultado um número de amostras consideralvemente menor que o inicial, sendo composto

Trabalho Analisado	Número de palavras catalogadas	Palavras com maior frequência de ocorrência	Palavras de maior conteúdo semântico
FPGA Architectures: Survey and Challenges	35521	Logic (446) FPGA (363) That (307) Routing (288) Block (243) This (242) Architecture (217) Area (189) Programmable (169) Design (154)	**Architecture** **FPGA** **Block**
A decade of Reconfigurable Computing: A Visionary Retrospective	9428	Data (67) Reconfigurable (65) With (59) Memory (54) Array (51) Routing (40) Time (46) Architecture (38) Compiler (38) Design (37) From (36) Mapping (34) Communication (33) Based (32) Computing (30)	**Reconfigurable** **Architecture** **Computing**
Reconfigurable Computing: A survey of Systems and Software	23523	Reconfigurable (334) Hardware (226) That (206) This (178) Computing (155) Logic (152) Configuration (120) FPGA (127) Routing (122) Circuit (118) FPGA's (116) with (109) these (104) Systems (103) Time (103) Programmable (91)	**Reconfigurable** **Hardware** **Computing** **FPGA**

Figura 5. Análise bibliometrica realizada pela ferramenta ZipfTool

de apenas trabalhos com a máxima convergência para com os objetivos da pesquisa.

Resultados Quantitativos

Após a redefinição dos termos de busca, a Figura 6 ilustra o número de trabalhos relacionados, os quais foram selecionados utilizando os veículos de publicação de maior ex-

pressão no campo de estudo conforme análise qualitativa descrita anteriormente e as palavras definidas pela ferramenta ZipfTool como parâmetros de busca.

Para ser possível a redução do universo de análise, além dos parâmetros identificados pela ferramenta ZipfTool juntamente a utilização dos veículos de publicações de maior expressão, foi necessário aplicar alguns conceitos estatísticos, principalmente no que tange à amostragem estratificada, pois, realizando uma análise mais minimalista pode-se observar que trabalhos de um mesmo ano e de um mesmo veiculo de publicação possuem características homogêneas entre si, como por exemplo a aplicação de conceitos, os quais são, objetos de estudo da pesquisa.

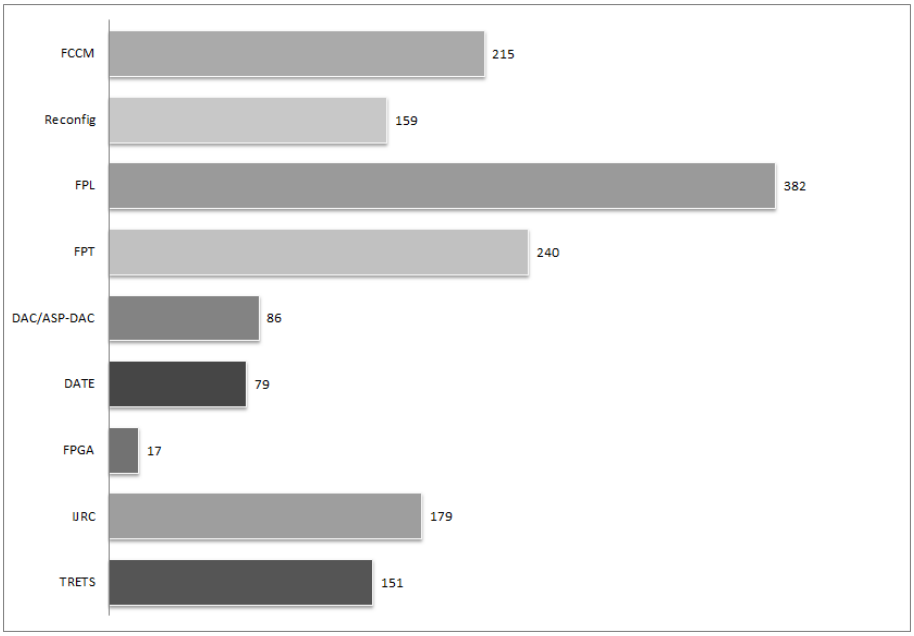


Figura 6. Número de Publicações nos principais veículos de divulgação

Através da Figura 6 pode-se observar uma grande redução do universo de análise. A ferramenta ZipfTool proporcionou a identificação dos principais termos que possuem relação direta os principais conceitos do campo de estudo, gerando assim, os parâmetros de busca necessário para seleção dos trabalhos impactantes para a pesquisa. Ainda, permitiu que os autores observassem características homogêneas entre determinados veículos de publicações, resultando na identificação de veículos mais expressivos para utilização da pesquisa, contribuindo diretamente para redução do universo amostral.

6 Conclusão

Neste trabalho foi apresentado o desenvolvimento da ferramenta ZipfTool, que trata-se de um mecanismo de análise de dados textuais com o principal objetivo de extrair do texto palavras de alto teor semântico, descritores e palavras chave. A ferramenta foi desenvolvida tendo como base os princípios da primeira e segunda Leis de Zipf's do campo da bibliometria, permitindo diversas opções de análise textual para o usuário.

A ZipfTool é uma ferramenta parametrizável, que possui um arquivo de configuração a partir dos qual é possível que o usuário controle diversos aspectos, tais como: caracteres e palavras de descarte, análise unitária ou em bloco e a geração de gráficos para visualização dos resultados obtidos.

Diante do atual cenário técnico/científico, onde os campos de estudos possuem inúmeras publicações em inúmeros veículos, uma ferramenta de auxílio que possibilita o usuário a ter uma melhor análise sobre os termos estudados, verificando a consistência dos termos utilizados bem como o nível de repetição de determinadas palavras, torna-se bastante importante.

Para ilustrar uma das funcionalidades da ferramenta, um estudo de caso foi apresentado, onde a ferramenta ZipfTool auxiliou os autores a entender de forma mais consistente e integra os conceitos que estavam observando, bem como, foi utilizada com filtro, auxiliando no descarte de trabalhos os quais não possuíam relação direta com os objetivos da pesquisa.

A análise feita sobre a ferramenta ZipfTool apresentada neste artigo considerou sua versão produzida sobre a plataforma MATLAB. Esta escolha deu-se pela facilidade de programação fornecida pelo MATLAB, uma vez que muitas primitivas matemáticas necessárias já encontram-se definidas em sua linguagem. Essa abordagem permitiu o desenvolvimento mais rápido de uma prova de conceito do que seria possível em outras linguagens. Entretanto, para que a ferramenta possa ser amplamente utilizada pela comunidade científica, temos consciência que a implementação de uma versão $on - line$ é necessária. Nesse sentido, nos propomos, à guisa de trabalho futuro, a programar de uma nova versão da ZipfTool baseada na linguagem $Python$, que será disponibilizada para uso através de um navegador de internet, sem necessidade de instalação no computador do usuário e que será gratuita.

7 *

Contribuição dos autores: - Diego Nunes Molinos: participou da elaboração do projeto, levantamento de requisitos, implementação e desenvolvimento da ferramenta, testes, análise de resultados, redação e revisão do artigo

- Daniel Gomes Mesquita: participou da elaboração do projeto e orientou todas as eta-

pas do trabalho incluindo o desenvolvimento da ferramenta, delineamento do estudo de caso, aprimoramento da revisão do estado-da-arte, redação e revisão final do artigo. Destaca-se a busca de referencial teórico que permitiu a adaptação da ZipfTool para o idioma português.

- Debora Nayar Hoff: contribuiu na definição dos argumentos relativos à construção do conhecimento e elementos constitutivos da revisão de literatura sobre bibliometria. Orientou todas as etapas de redação e revisão do artigo. Destaca-se neste processo sugestões de melhoria das descrições das experiências, análises e formação de quadros resumo dos resultados.

8 Referências

9 *

Referências

[1] N. Abbagnano. Tradução: Alfredo bosi - dicionário de filosofia. *Dicionário de filosofia*, 2, 1970.

[2] A. D. Booth. A "law"of occurrences for words of low frequency. *Information and control*, 10(4):386–393, 1967.

[3] P. Cabena, H. H. Choi, I. S. Kim, S. Otsuka, J. Reinschmidt, and G. Saarenvirta. Intelligent miner for data applications guide. *IBM RedBook SG24-5252-00*, 173, 1999.

[4] B. Camargo and A. Justo. Tutorial para uso do software de análise textual iramuteq. *Florianopolis-SC: Universidade Federal de Santa Catarina*, 2013.

[5] C. J. G. Campos. Método de análise de conteúdo: ferramenta para a análise de dados qualitativos no campo da saúde. *Rev Bras Enferm*, 57(5):611–4, 2004.

[6] Y.-S. Chen and F. F. Leimkuhler. Analysis of zipf's law: An index approach. *Information processing & management*, 23(3):171–182, 1987.

[7] K. Compton and S. Hauck. Reconfigurable computing: a survey of systems and software. *ACM Computing Surveys (csuR)*, 34(2):171–210, 2002.

[8] I. Dahlberg. Teoria do conceito. *Ciência da informação*, 7(2), 1978.

[9] S. de Amo. Técnicas de mineração de dados. *Jornada de Atualizaçao em Informatica*, 2004.

[10] R. Feldman and J. Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.

[11] A. Gilat. *MATLAB com aplicações em Engenharia*. Bookman, 2006.

[12] W. Goffman and V. Newill. Generalization of epidemic theory. *Nature*, 204(4955):225–228, 1964.

[13] V. L. Guedes and S. Borschiver. Bibliometria: uma ferramenta estatá■stica para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação cientí■fica e tecnologica. *Encontro Nacional de Ciência da Informação*, 6:1–18, 2005.

[14] R. Hartenstein. A decade of reconfigurable computing: a visionary retrospective. In *Proceedings of the conference on Design, automation and test in Europe*, pages 642–649. IEEE Press, 2001.

[15] IEEE. ieeexplore digital library. URL: http://www.ieeexplore.ieee.org/Xplore/home.jsp, 2013. Acesso em 17/05/2013.

[16] T. S. Jean. An introduction to informetrics. *Information processing management*, 28(1):1–3, 1992.

[17] M. Klemann, E. Reategui, and C. Rapkiewicz. Análise de ferramentas de mineração de textos para apoio a produção textual. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 1, 2011.

[18] D. Kuhn. Teaching and learning science as argument. *Science Education*, 94(5):810–824, 2010.

[19] E. Lima and S. Maia. Comportamento bibliométrico da língua portuguesa, como veículo de representação da informação. *Ciência da Informação*, 2(2), 1973.

[20] S. S. Lisboa. A importancia dos conceitos da geografia para a aprendizagem de conteúdos geográficos escolares. *CEP*, 36570:000, 2007.

[21] A. L. Macedo, E. Reategui, A. Lorenzatti, and P. Behar. Using text-mining to support the evaluation of texts produced collaboratively. In *Education and Technology for a better world*, pages 368–377. Springer, 2009.

[22] G. d. A. Martins. Sobre conceitos. definições e constructos nas ciências administrativas. *Gestão & Regionalidade*, 21(62), 2010.

[23] D. N. Molinos. Arcabouço conceitual para computação reconfigurável. URL: http://http://repositorio.ufu.br/handle/123456789/4550, 2014. Acesso em 03/01/2015.

[24] M. Z. Moretto and C. E. Rapkiewicz. Usando mineração de textos como suporte ao desenvolvimento de resumos no ensino médio. *RENOTE*, 11(3), 2013.

[25] S. Morgan Friedman. Wordcounter. URL: http://http://www.wordcounter.com, 2004. Acesso em 17/06/2014.

[26] P. Ratinaud. Iramuteq: Interface de r pour les analyses multidimensionnelles de textes et de questionnaires. *Téléchargeable à l?adresse: http://www.iramuteq.org*, 2009.

[27] J. Rose, I. Kuon, and R. Tessier. Fpga architecture: Survey and challenges. *Foundations and Trends® in Electronic Design Automation*, 2(2):135–253, 2008.

[28] M. d. S. Soares. Comparação entre metodologias áĄgeis e tradicionais para o desenvolvimento de software. *INFOCOMP Journal of Computer Science*, 3(2):8–13, 2004.

[29] D. Steinbock. Tagcrowd. URL: http://http://tagcrowd.com/, 2002. Acesso em 17/06/2014.

[30] J. H. Y. F. W. Wang, J. C. W. G. K. Koperski, D. Li, Y. L. A. R. N. Stefanovic, and B. X. O. R. Zaiane. Dbminer: A system for mining knowledge in large relational databases. In *Proc. Intl. Conf. on Data Mining and Knowledge Discovery (KDD?96)*, pages 250–255, 1996.

[31] V. R. Werneck. Sobre o processo de construção do conhecimento: o papel do ensino e da pesquisa. *Ensaio: Avaliação e Políticas Públicas em Educação*, 14(51):173–196, 2006.

[32] R. K. Yin. *Estudo de caso: Planejamento e métodos*, volume 4. Bookman Porto Alegre, 2005.

[33] G. K. Zipf. Relative frequency as a determinant of phonetic change. *Harvard studies in classical philology*, pages 1–95, 1929.

Escola Gaúcha de Bioinformática

27 a 31 de Julho de 2015
Centro de Eventos do Instituto de Informática da UFRGS

A Bioinformática é um dos campos de pesquisas com maior crescimento nos últimos anos e cujo papel nas diferentes áreas do conhecimento biológico tem aumentado de forma exponencial. Este crescimento se deve a inúmeros fatores, dos quais se destacam: (i) geração de dados biológicos em grandes volumes, provenientes do desenvolvimento das chamadas "técnicas de larga-escala", das quais incluem-se a genômica, transcriptômica e proteômica; (ii) acesso facilitado a bancos de dados desenvolvidos especialmente para a estocagem dos dados biológicos; (iii) geração de algoritmos e sistemas informatizados capazes de processar os diferentes dados biológicos e (iv) diminuição dos custos associados com processadores, memórias e sistemas de armazenamento, facilitando o acesso de diferentes grupos de pesquisas à computadores com alto poder e/ou desempenho para o processamento de dados. Conectado a estes pontos está o fato de que a Bioinformática, como campo de pesquisas, alia de forma transdisciplinar as áreas das Ciências Biológicas e das Exatas (Ciência da Computação, Física, Matemática, entre outras), proporcionando uma troca bastante dinâmica de conhecimentos entre os pesquisadores.

A Escola Gaúcha de Bioinformática, EGB, realizada de 27 a 31 de julho de 2015, foi organizada de forma integrada pelo Instituto de Informática e pelo Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul, buscando construir um espaço de formação, integração, qualificação e desenvolvimento das atividades de pesquisa envolvendo o emprego de métodos computacionais no estudo de sistemas biológicos. Sob uma perspectiva interdisciplinar de problemas biológicos e terapêuticos, o evento agregou tanto abordagens através de sequências (de nucleotídeos ou aminoácidos), de biologia de sistemas e de larga escala, quanto de estruturas 3D (e suas conformações) ao suporte e desenvolvimento possibilitados pela Ciência da Computação. Durante a realização da EGB foram ofertadas aulas, mini-cursos e atividades envolvendo assuntos do estado da arte da área de Bioinformática, os quais não são frequentemente cobertos nos cursos graduação e ou pós-graduação no estado do RS. Além disto, o evento induziu futuras colaborações entre alunos e pesquisadores de diferentes instituições de ensino e pesquisa do Rio Grande do Sul e Brasil. A Escola também contribuiu com a qualificação de profissionais que atuam na área de Bioinformática no estado do Rio Grande do Sul e também com o desenvolvimento e expansão desta área de pesquisa. A EGB 2015 contou com a participação efetiva de 159 pessoas. Durante a semana de realização da Escola estiveram presentes estudantes de graduação e pós-graduação, bem como de profissionais das áreas de Ciências Biológicas, Matemática, Física, Química, Engenharias e Ciências da Computação e as suas respectivas áreas correlatas.

Durante a EGB foram realizados 7 mini-cursos teóricos, práticos e teórico-práticos em nível básico e avançado, cada um com carga horária de 15hs. Um público de 133 pessoas participaram dos mini-cursos no período 27 a 31 de Julho de 2015. Os mini-cursos realizados e o número de participantes em cada mini-curso foram: Mini-curso 01: Análises transcriptômicas e biologia de sistemas - avançado: 31 participantes; Mini-curso 02: Programação Java para Bioinformática - básico: 11 participantes; Mini-curso 03: Introdução à Biologia de Sistemas - básico: 36 participantes; Mini-curso 04: Análise e interpretação de dados estruturais e conformacionais - avançado: 9 participantes; Mini-curso 05: Programação Python para Bioinformática - básico: 30 participantes; Mini-curso 06: Uso de pacote R para a Biologia de Sistemas - básico: 8 participantes; Mini-curso 07: Biologia Estrutural para iniciantes - básico: 8 participantes. Durante a EGB 2015 foram aceitos e apresentados, na forma de pôsteres e sessões colaborativas, 42 trabalhos científicos. Os trabalhos foram submetidos na forma de resumos, e encontram-se nas próximas páginas.

A Comissão Organizadora agradece: a presença de estudantes, professores, pesquisadores e palestrantes; ao apoio financeiro recebido pela FAPERGS, CNPq, CAPES e PROPESQ; ao Instituto de Informática da UFRGS e sua equipe técnica pelo suporte recebido durante o evento; e aos estudantes do Instituto de Informática e do Centro de Biotecnolgia da UFRGS que estiveram envolvidos com a organização da EGB 2015.

Porto Alegre, 30 de Maio de 2016.

Márcio Dorn – Inf – UFRGS
Diego Bonatto – CBiot – UFRGS
Hugo Verli – CBiot - UFRGS

Identification of Transcription Factors that Act as Master Regulators in Alzheimer's Disease and Parkinson's Disease

Marco Antônio De Bastiani[1]
Carolina Chatain[1]
Mauro Antônio Castro[2]
Fábio Klamt[1]

Alzheimer's Disease (AD) and Parkinson's Disease (PD) are the two most common neurodegenerative disorders. It is estimated that more than 45 million people worldwide suffer from one of these pathologies. Despite the large investment in the neuroscience field, etiology and molecular mechanisms underlying neuronal death remain unclear. Recently, a small proportion of cases of AD and PD have been attributed to mutations in specific genes. However, the many pathways in which their gene products are involved and the interaction with other factors that might lead to neuropathological changes are still poorly understood. In the current work, microarray data acquired from NCBI public repository Gene Expression Omnibus (GEO) (GSE60862) was used to determine normal tissue-specific transcriptional networks for hippocampus and *substantia nigra*, structures specifically damaged in AD and PD, respectively. Case-control studies (GSE5281 and GSE8397), also obtained from GEO, were used to establish gene expression signatures for the two neurodegenerative diseases and identify transcription factors (TFs) that are pivotal to modulate phenotypic changes from normal to pathological contexts, called master regulators (MRs), applying MRA and GSEA algorithms. As results, we identified 117 important TFs regulating gene expression in hippocampus and 123 in *substantia nigra*. We proposed 17 MRs involved with AD and 28 with PD, some of which have already been described in the process of neurodegeneration (such as YY1, HMG20A, RREB-1and SLC3A9) and others not related to the pathologies so far. We believe that these results might help in the understanding of AD and PD and lead to the discovery of targets for potential therapeutic intervention.

[1] Universidade Federal do Rio Grande do Sul
{marco.bastiani@ufrgs.br, carolpchatain@gmail.com, 00025267@ufrgs.br}
[2] Universidade Federal do Paraná {mauro.a.castro@gmail.com}

GROMOS 53A6 Force Field Parameters for Chalcones

Elisa Beatriz de Oliveira John[1]
Pablo Ricardo Arantes[1]
Hugo Verli[1]

Chalcones are compounds extensively distributed in plants and considered as the precursors for flavonoid synthesis. Their structure consists of an open chain in which two aromatic rings are joined by a three-carbon α, β-unsaturated carbonyl system. Chalcones are classified according to the A- and B- ring substitution by OH and OCH_3 groups, and different substitution patterns can affect their biological activities. Since they present many pharmacological activities, these biomolecules are being intensively studied and modified. Computational methods such as molecular dynamics (MD) simulations can provide microscopic information that may be difficult to obtain from other experiments. Accurate force fields are essential for describing biological systems in a MD simulation, thus a parameter set associated to a certain compound need to be carefully calibrated to ensure reliable results. There are no validated parameters for the α, β-unsaturated carbonyl group in the GROMOS force field, therefore the present work intends to provide a new parameter set for the simulation of chalcones. We employed a protocol for parameter optimization combining ab initio calculations, which provide the quantum mechanical (QM) potential energy profile (performed by GAMESS), and MD simulations that provide the molecular-mechanical (MM) potential energy profile, using the GROMACS simulation suite and GROMOS53A6 force field. A fitting of MM to QM torsional profiles was performed for each of three dihedrals of interest within the basic structure of chalcone. For the dihedrals including the rings, the fitting has already generated parameters that reproduce well the QM potential energy in the MM calculations. Additionally, the density and enthalpy of vaporization values are in good agreement with experimental data. When completed, we expect that such parameters will be able to properly describe the conformational distribution of chalcones, a starting point to further studies on the biological role of such molecules, at the atomic level.

[1] Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Caixa Postal 15005
{elisajohn@live.com; hverli@cbiot.ufrgs.br}

Effects of D-myo-inositol 3,4,5,6-tetrakisphosphate (TMI) binding on antithrombin

Pablo Ricardo Arantes[1]
Horacio Pérez-Sánchez[2]
Hugo Verli[1]

Antithrombin (AT), a serine protease inhibitor, circulates in blood in two major isoforms, α and β, which differ in their amount of glycosylation and affinity for heparin. After binding to this glycosaminoglycan, the native AT conformation, relatively inactive as a protease inhibitor, is converted to an activated form. Recently a new compound, named TMI, was discovered with nanomolar affinity to antithrombin and shown to be able to induce a partial activation of antithrombin, which in turn facilitates the interaction with heparin. In this context, the present work intends to characterize the effects of TMI binding on AT structure and dynamics through a series of MD simulations. The GROMACS package was employed with GROMOS 43a1 force field, the native and activate α-AT unbounded and bounded to heparin and TMI were selected for simulations under SPC water models, the PME method was applied in the calculation of electrostatic interactions, and simulations were performed in triplicate at 310K for 0.2μs. Molecular mechanics Poisson - Boltzmann surface area (MMPBSA) was used to estimate binding free energy of heparin and TMI, showing a correlation to the experimentally determined affinities. The reactive center loop (RCL), between residues Glu377 to Leu400, on the TMI-bounded AT simulations shown a flexibility behavior similar to the observed for heparin-bounded AT. As well, TMI also exposes the P1 residue of antithrombin (Arg393), as observed during the heparin-bounded AT. Combined, these data provides atomic details for TMI induced partial activation of AT, and may constitute a basis for future studies aiming TMI structural optimization.

Supported by: CNPq, CAPES and FAPERGS

[1]Centro de Biotecnologia, UFRGS
[2] Bioinformatics and High Performance Computing Research Group, Universidad Católica San Antonio de Murcia (UCAM), Spain
{pabloarantes@cbiot.ufrgs.br; hperez@ucam.edu; hverli@cbiot.ufrgs.br}

The complete chloroplast genome sequence of neotropical Myrtaceae Eugenia uniflora: organization and phylogenetic relationships

Eguiluz M. Maria[1]
Rodrigues F. Nureyev[1]
Guzman Frank[1]
Yuyama Priscila[2]
Margis Rogerio[123]

Eugenia uniflora is plant native to tropical America with pharmacological and ecological importance. The complete chloroplast (cp) genome sequence of Eugenia uniflora, the first sequenced member of the neotropical myrtaceae family, is reported here. The genome is 158,445 bp in length and exhibits a typical quadripartite structure of the large (LSC, 87 459 bp) and small (SSC, 18 318 bp) single-copy regions, separated by a pair of inverted repeats (IRs, 26 3334 bp). It contains 111 unique genes, including 77 protein-coding genes, 30 tRNAs and four rRNAs. The genome structure, gene order, GC content and codon usage are similar to the typical angiosperm cp genomes. Entire cp genome comparison of E.uniflora and three others Myrtaceas revealed an expansion in the intergenic spacer located between IRA/large single copy (LSC) border and the first gene of LSC region, driven by sequence of 43 bp. Simple sequence repeat (SSR) analysis revealed that most SSRs are AT-rich, which contribute to the overall AT richness of the cp genome. Additionally, fewer SSRs are distributed in the protein-coding sequences compared to the noncoding regions. Phylogenetic analysis demonstrated a close relationship between Eugenia uniflora and Syzygium cuminis in Myrtaceae. The complete cp genome sequence of E.uniflora reported here will facilitate population, phylogenetic and evolutionary studies

[1] PPGGBM, Departamento de Genética, Universidade Federal do Rio Grande do Sul - UFRGS,Porto Alegre, Rio Grande do Sul, Brazil
[2] Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil
[3] Departamento de Biofisica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil

Insights on the Dynamics and Thermostability of an Archaeal Oligosaccharyltransferase

Conrado Pedebos[1]
Hugo Verli[1]

N-glycosylation is one of the most prevalent post-translational modifications. The addition of newly formed glycan chains to a nascent polypeptide is fulfilled by the oligosaccharyltransferases PglB (Bacteria), AglB (Archaea), and Stt3 (catalytic subunit in Eukarya). Previous crystallographic data for these enzymes identified structural units with distinct functions, such as catalysis (CC) and structural stability (IS, P1, and P2). Studies regarding these enzymes and its units may support the development of vaccines and glycoprotein engineering. Therefore, here, we examine the predicted role of P1 unit from AglB, aiming to contribute with insights for the comprehension of AglB thermostability. We performed molecular dynamics simulations of the following systems: i) wild type AglB, ii) AglB lacking the P1 unit (AglBΔP1), and iii) P1 unit, as a control. The force field employed was GROMOS54a7, at a temperature of 356 K, in the presence of explicit aqueous solvent, catalytic ion, and a membrane bilayer composed of palmitoyl-oleoyl-phosphatidylethanolamine lipids. AglB maintained a stable behavior during simulations, while the AglBΔP1 structure became highly unstable, losing contacts and secondary structure elements, although not influencing the large transmembrane section. Nevertheless, AglBΔP1 maintained some organization at the catalytic site, such as the coordination of the metal ion with the catalytic residues, which indicates the many roles played by Zn2+, acting both as a catalytic and as a structural ion. Additionally,we detected regions with conserved residues that preserves the strong interaction between P1 unit and AglB, and could be the target of mutational studies. This data may provide the basis for the engineering of thermostable oligosaccharyltransferases from different species. The insertion of a P1 unit at the C-terminal end of the bacterial PglB would be the first step on the generation of chimeric OSTs with biotechnological applications.

Supported by: CNPq, CAPES and FAPERGS

[1] Centro de Biotecnologia, UFRGS, Caixa Postal 15005
{conrado.pedebos@ufrgs.br; hverli@cbiot.ufrgs.br}

A comparison of molecular dynamics simulations using GROMACS with GPU and CPU

Alex Dias Camargo[1],
Adriano V. Werhli[1],
Karina dos Santos Machado[1]

1 Introduction

According to Zhou et al. [13] mathematical modelling is central to systems and synthetic biology . The use of simulations is a common means for analysing these models and it is, usually, computationally intensive. High-performance computing holds the key to making significant biomolecular calculations [4] and the complexity of computational calculations has historically been extremely high [3]. The Molecular Dynamics (MD) simulations share a need for fast and efficient software that can be deployed on massive scale in distributed computing [7]. MD is based on Newton's perception, that from the starting position, it is possible to calculate the next position and velocities of the atoms in a small time interval and the forces in the new position [2]. MD is used in physics, biology and chemistry where systems of several million of atoms are simulated for days or weeks prior to completion [12]. Hardware advances brings supercomputing power to the desktop computer, thus facilitating the widespread use of parallel algorithms by bioinformaticians [11]. GPUs (Graphics Processing Units) are different from CPUs (Central Processing Units) in several fundamental ways that impact how they can be executed [10]. Various MD implementations that utilize GPUs to gain notable performance over CPUs have been described [1, 8, 9]. In this paper we show the performance of MD simulations using versions of GROMACS (Groningen Machine for Chemical Simulation) on GPU and CPU. The GPU performance was evaluated and it is shown that it can be more than 50% faster than a conventional method running on a multiple CPU core. This paper is organized as follows: Important features of the GPU architecture and CUDA (Compute Unified Device Architecture) programming model are described in Section 2. Section 3 introduces the MD and the software suite GROMACS. The obtained results are described in Section 4. Finally, Section 5 presents the discussion and conclusion.

2 GPU architecture and CUDA

A GPU is a massively parallel computer designed to accelerate computationally intensive applications which operate in a single-instruction multiple-thread (SIMT) mode [11]. With the enhanced programmability of GPUs, these chips are now capable of performing more than the specific graphics computations they were originally designed [4]. CUDA Toolkit provides a comprehensive development environment for developers building GPU-accelerated applications. It includes a compiler for NVIDIA GPUs, math libraries and tools for optimizing the performance of applications [6]. We have implemented the proposed application using CUDA Toolkit 7.

[1]Programa de Pós-graduação em Computação, FURG, Caixa Postal 96201-900
{alexcamargo, werhli, karina.machado}@furg.br

3 Molecular Dynamics Simulation and GROMACS

The MD emerged as one of the first simulation methods from the pioneering applications to the dynamics of liquids by Alder and Wainwright and by Rahman in the late 1950s and early 1960s [5]. Modeling and visualization of atomic level details provides insight into the function and dynamics of biomolecular structures [10]. The core of process is the computation of distances between pairs of atoms, and subsequently the forces and/or energies that they exert on each other [12]. In accordance with [3], a good MD simulation in GROMACS depends on three factors: the speed of the computational part in isolation, the efficiency of the parallel and communication algorithms, and the efficiency of the communication itself. In this paper we use GROMACS[1] version 5, since this software has native support for GPUs, in comparison by version 4 (OpenMP trheads native). Therefore, it is possible to perform high-throughput MD simulations. Algorithm 1 [12] illustrates the pseudocode for a MD simulation on GPU.

Algorithm 1. Pseudocode for GPU MD simulations [12]

```
1: Initialize the position, velocity, and force of atoms
2: Allocate memory for the atoms on the GPU
3: Transfer force, velocity, and position data to GPU
4: for all steps do
5:   Launch position GPU kernel
6:        Update head atom positions
7:        Half-update head atom velocity basedon previous force
8:        Zero head atom forces
9:        Launch force GPU kernel
10:  for all atoms do
11:    Compute force exerted by atom on head atom if within cutoff distance
12:  end for
13: Complete update of head atom velocity based on current force
14: end for
15: Transfer data from GPU to CPU
```

4 Performance Results

This paper compares the use of the software suite GROMACS running entirely on a GPU and CPU. CPUs typically provide a small number of very fast processing units, whereas GPUs have a large number of slower processing units [3]. We considered as a case study the main steps of a MD simulation of a system containing the protein Lysozyme (PDB code 1AKI) in a box of water having 38,790 atoms including solvent and ions for general purposes.

The sequential MD simulations tests were executed on identical hardware on a PC with an Intel Xeon CPU X5675 – 3,6GHz (12 processor cores), RAM 12GB, HD 1TB, running Ubuntu 14.04 64 bit Desktop Edition. A single GPU NVIDIA Quadro 600 (96 CUDA cores) was used. Table 1 shows the performance comparison between the two run methods, GROMACS 4 on CPU and GROMACS 5 on GPU, for main tasks of MD.

[1] http://www.gromacs.org/downloads

Table 1. Comparison of runtimes

TASK	Gromacs 4 (s)	Gromacs 5 (s)
Energy minimization	31,77	16,27
Equilibration (phase 1)	6207,78	2188,51
Equilibration (phase 2)	5960,86	2238,95
Production MD	22766,65	18267,44

In energy minimization, the structure is relaxed avoiding steric clashes or wrong geometry. Maximum number of steps (nsteps) to perform was 5000. This task demands less throughput data. The GROMACS 5 gain over GROMACS 4 total time reached 49%. In equilibration (phase 1 and phase 2), the structure is to brought to simulation temperature and establish the ideal orientation about the solute, also it is needed to stabilize the pressure of the system. In lines 3 and 4 of Table 1 it can be seen that the GROMACS 5 performance over GROMACS 4 total time was respectively 65% and 62% faster respectively, with nsteps = 5000. Upon completion of the two equilibration phases, we are ready to run production MD for data collection. With nsteps = 500000, GROMACS 5 perfomance also exceeded by 20% of the GROMACS 4 total time. These results show the advantages of GPU use whenever we have a high throughput data.

5 Discussion and conclusion

MD simulations of macromolecules are extremely computationally demanding, which makes them a natural candidate for implementation on GPUs [10]. The time required for such a step depends normally on the system. Advances in performance make it possible to calculate complex biomolecular interactions and function using a single desktop computer.

In this paper we disucss how MD simulations can benefit from the computing power of GPUs. The results show that the performance of MD simulations of proteins on GPU using GROMACS is attractive. As future work we intend to run MD simulations with different proteins and MD parameters, also considering other GPUs.

Acknowledgment

This work was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

References

[1] Anderson, Joshua A., Chris D. Lorenz, and Alex Travesset. "General purpose molecular dynamics simulations fully implemented on graphics processing units." Journal of Computational Physics 227.10 (2008): 5342-5359.

[2] Astuti, A. D., and A. B. Mutiara. "Performance Analysis on Molecular Dynamics Simulation of Protein Using GROMACS." arXiv preprint arXiv:0912.0893 (2009).

[3] Hess, Berk, et al. "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation." Journal of chemical theory and computation 4.3 (2008): 435-447.

[4] Liu, Weiguo, et al. "Molecular dynamics simulations on commodity GPUs with CUDA." High Performance Computing–HiPC 2007. Springer Berlin Heidelberg, 2007. 185-196.

[5] Meller, Jaro. "Molecular dynamics." ELS (2001).

[6] NVIDIA, CUDA. "CUDA Toolkit Documentation - v7.0 ". Available: https://docs.nvidia.com/cuda/. (2015).

[7] Pronk, Sander, et al. "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit." Bioinformatics (2013): btt055.

[8] Schmid, Nathan, Mathias Bötschi, and Wilfred F. Van Gunsteren. "A GPU solvent–solvent interaction calculation accelerator for biomolecular simulations using the GROMOS software." Journal of computational chemistry 31.8 (2010): 1636-1643.

[9] Stone, John E., et al. "Accelerating molecular modeling applications with graphics processors." Journal of computational chemistry 28.16 (2007): 2618-2640.

[10] Rodrigues, Christopher I., et al. "GPU acceleration of cutoff pair potentials for molecular modeling applications." Proceedings of the 5th conference on Computing frontiers. ACM, (2008).

[11] Vouzis, Panagiotis D., and Nikolaos V. Sahinidis. "GPU-BLAST: using graphics processors to accelerate protein sequence alignment." Bioinformatics 27.2 (2011): 182-188.

[12] Walters, John Paul, et al. "Accelerating Molecular Dynamics Simulations with GPUs." ISCA PDCCS. (2008).

[13] Zhou, Yanxiang, et al. "GPU accelerated biochemical network simulation." Bioinformatics 27.6 (2011): 874-876.

Dynamics of the Complete, Membrane Soaked hTLR4

Carla Carvalho de Aguiar[1]
Hugo Verli[1]

Human Toll-Like Receptor 4 (hTLR4) is a transmembrane protein of the immune system. Together with MD-2, a co-receptor, hTLR4 recognizes bacterial lipopolysaccharide. hTLR4 posses three domains: an extracellular domain (ECD; for ligand recognition); an helical transmembrane domain (TM); and an intracellular domain (TIR; for signaling). In a previous work, our group described the influence of the co-receptor over the hTLR4 ECD dynamics. As this model included only the receptor ECD, the present work aims to study the relation of hTLR4 and MD-2 within membranes, searching for clues to the receptor conformational activation. For this, comparative modeling (Modeller v.9) was employed to produce a model of the complete hTLR4, inserted in a POPC membrane as a single (i, hTLR4) and heterodimeric (ii, hTLR4-MD-2) systems. Both i and ii were submitted to molecular dynamics simulations using GROMACS and GROMOS 53a6 force field. We observed that MD-2 was able to interfere in hTLR4 conformational space, whereas it does not seems to be related to the principal movement of ECD. In both systems, the TIR and ECD became closer to the membrane, and Leu661 of TM seems to be involved in the conformational connection between TM and TIR. Moreover, dynamical network analysis shows that MD-2 induced the formation of a more adherent community of conformational connected residues in TIR. Once we had inferred, previously, a main movement of ECD in solution, now we could identified this behavior without the bias of the absence of other domains or membrane. Furthermore, TM mobility could be influencing TIR dynamics in relation to membrane through specific residues.

Supported by: CNPq, CAPES and FAPERGS

[1]Centro de Biotecnologia, UFRGS, Caixa Postal 15005
{carlaaguiar@cbiot.ufrgs.br; hverli@cbiot.ufrgs.br}

A proposal to the manipulatioon of a set of protein structures from PDB

Vinicius Rosa Seus [1]
Karina dos Santos Machado[2]
Adriano Velasque Werhli[3]

Protein Data Bank (PDB) is a public web database with more than 100,000 biological macromolecular structures. With this large amount of protein structures available on PDB the use of tools for acquisition and analysis of specific sets of biological macromolecules is a necessity. Hence, in this work we propose the development of a tool for acquiring, storing and analyzing specific sets of proteins from PDB. The proposed tool runs on desktop environment allowing the user to acquire the structures from the RESTful web-service provided by PDB server. After the acquisition of a set of interesting PDBs the user can manipulate these data in an off-line environment through a local database that stores the information about the characteristics of the structures, for example, ligands, mutations, residues, sequences and docking results. The protein files are locally stored on users' computer and can be used, for instance, for molecular docking simulations and alignment of sequences and structures. Having a set of proteins of interest available locally and using our proposed tool the user can perform analysis related to alignments and visualize important proteins characteristics improving the knowledge about specific target. Besides, the user can select PDB files to be visualized on a graphical environment that is integrated in our tool. Other features are related to the exporting of sequence alignments results in csv format or exporting sequences that have a similar identity in a format that can be easily loaded on graph tools. These alignments allow user to visualize which proteins are similar and discard those that are not. As future work we propose to conduct an study using our tool for acquiring and analyzing a set of betaglucosidase molecules, which is the enzyme responsible for the extraction of fermentable sugars from the sugar cane bagasse used in the industry for bio-ethanol production.

[1] C3 - Centro de Ciências Computacionais, FURG, Av. Itália km 8 Bairro Carreiros, Brazil, Rio Grande do Sul, Rio Grande {viniciusseus@furg.br}
[2] C3 - Centro de Ciências Computacionais, FURG, Av. Itália km 8 Bairro Carreiros, Brazil, Rio Grande do Sul, Rio Grande {karina.machado@furg.br}
[3] C3 - Centro de Ciências Computacionais, FURG, Av. Itália km 8 Bairro Carreiros, Brazil, Rio Grande do Sul, Rio Grande {werhli@furg.br}

Mother´s Vaginal Microbiota Shares Few Phylotypes With Preterm Infants

Priscila Caroline Thiago Dobbler[1]
Miriane Acosta Saraiva[2]
Andréa Lúcia Corso[2]
Rita de Cassia Silveira[2]
Renato Soibelmann Procianoy[2]
Luiz Fernando Wurdig Roesch[1]

Evidences generated with new cultivation free molecular tools, challenges the established dogma that the fetus remains sterile until delivery. Microbes derived from the mother's vagina, gut or placenta might colonize the fetus's gastrointestinal tract, however, the individual contribution of each mother's body site to such early colonization is still unknown. Here we determined the percentage of microbes shared between the mother's vagina and the newborn. For this analysis, we performed microbial DNA extraction of a single vaginal swab from 11 pregnant women and second evacuation of their respective preterm infants. The gestational period ranged from 27 to 32 weeks. Nine mothers gave birth through C-section and two had normal (vaginal) delivery. One mother gave birth to twins. Following DNA extraction, the V3 region of the 16S rRNA gene was amplified from all samples, using the 515F and 806R primers and the amplicons were sequenced using the Ion Torrent-PGM technology. Quality filtering and assembling of sequences into Operational Taxonomic Units were performed according to the Brazilian Microbiome Project pipeline, available at: http://www.brmicrobiome.org. Microbial phylotypes shared between mother´s vagina and preterm infants varied among samples. Mothers shared 0 to 42.59% (Average of 13.68% and SD = 13%) of their vaginal microbiota with their infants. Moreover, the newborn's microbiome comprised of around 0 to 34% (Average of 15% and SD = 10%) of phylotypes that were similar to the mother's vagina. In summary, our results indicated that the microbiota from mother's vagina is a source of colonization of the fetus's gastrointestinal tract but the number of microbes shared between mother and the fetus is variable and in some cases, no microbes derived from the mother's vagina. These results indicated that the fetus might also receive microbes from another source, such as the placenta or the mother´s gastrointestinal tract.

[1] Universidade Federal do Pampa – São Gabriel, UNIPAMPA, Endereço: Av. Antônio Trilha, 1847 - São Gabriel - RS - CEP: 97300-000
{Priscila Caroline Thiago Dobbler, prisciladobbler@gmail.com} {Miriane Acosta Saraiva, mirisdabio@gmail.com} {Luiz Fernando Wurdig Roesch, luizroesch@unipampa.edu.br}
[2] Universidade Federal do Rio Grande do Sul, UFRGS, Av. Paulo Gama, 110, Porto Alegre
{Andréa Lúcia Corso, andrea.lucia@terra.com.br} {Rita de Cassia Silveira, drarita.c.s@gmail.com}

CrossTope: how this database can impact new vaccine strategies development

FREITAS, Martiela Vaz[1]
BRAGATTE, Marcelo Alves[1]
ANTUNES, Dinler Amaral[1,2]
RIGO, Maurício Menegatti[1]
MENDES, Marcus de Almeida[1]
SINIGAGLIA, Marialva[1]
VIEIRA, Gustavo Fioravanti[1]

One of the greatest challenges in immunology is the discovery of appropriate targets to vaccine development. The choice relies on the identification of elements responsible for the stimulation of immune responses. These features could define the cross-reactivity among different pMHC-I complexes (one T Cell Receptor-TCR could recognize more than one pMHC-I). Focusing on cross-reactivity, our group developed the CrossTope Data Bank, a curated repository of 3D structures of pMHC-I complexes.]One of the greatest challenges in immunology is the discovery of appropriate targets to vaccine development. The choice relies on the identification of elements responsible for the stimulation of immune responses. These features could define the cross-reactivity among different pMHC-I complexes (one T Cell Receptor-TCR could recognize more than one pMHC-I). Focusing on cross-reactivity, our group developed the CrossTope Data Bank, a curated repository of 3D structures of pMHC-I complexes. One of the greatest challenges in immunology is the discovery of appropriate targets to vaccine development. The choice relies on the identification of elements responsible for the stimulation of immune responses. These features could define the cross-reactivity among different pMHC-I complexes (one T Cell Receptor-TCR could recognize more than one pMHC-I). Focusing on cross-reactivity, our group developed the CrossTope Data Bank,a curated repository of 3D structures of pMHC-I complexes. One of the greatest challenges in immunology is the discovery of appropriate targets to vaccine development. The choice relies on the identification of elements responsible for the stimulation of immune responses. These features could define the cross-reactivity among different pMHC-I complexes (one T Cell Receptor-TCR could recognize more than one pMHC-I). Focusing on cross-reactivity, our group developed the CrossTope Data Bank, a curated repository of 3D structures of pMHC-I complexes.

[1]Núcleo de Bioinformática do Laboratório de Imunogenética, Departamento de Genética, Universidade Federal do Rio Grande do Sul, Postcode 91501-970, Brazil; {imunoinfo@gmail.com}

[2]Department of Computer Science, Rice University, Houston, Texas, 77005, USA. {dinler@gmail.com}

The Proteins are Linearized before the Proteasome Degradation: Are you sure?

TARABINI, Renata Fioravanti[1]
GUTIERRES, Matheus de Bastos Balbé[1]
FREITAS, Martiela Vaz[1]
SINIGAGLIA, Marialva[1]
VIEIRA, Gustavo Fioravanti[1]

The proteasome is a protein complex composed by a central pore with proteolytic activity combined with two regulatory components that recognize polyubiquitinated proteins and direct them to the catalytic core for degradation. This pathway is the main mechanism of protein digestion in eukaryotes and presents an important role in processing and presentation of epitopes to T cells in cell-mediated immunity. It is well known that the substrate enters the proteasome central cavity lacking its native 3D-structure, and then passes to the substrate binding channel where the peptide bond hydrolysis takes place. Research has focused efforts manly on the substrate primary sequence influence to the substrate processing inside the channel and still remains dubious the existence of some level of secondary structure. This fact suggest the hypothesis that spatial characteristics among the spectrum of secondary structure (beta sheets and alpha helices), could drive preferential cleavage sites. Our group has been working on a way to integrate and automate the interpretation of the outputs from Netchop 3.1 and Stride programs. Given a pdb file as an input for our workflow, Stride program assigns secondary structure for each amino acid (AA) and generate files in fasta format, which are used as input to the Netchop 3.1 program. In this program, the cleavage sites are predicted and recovered when an associated probability (higher than 70%) is found. The script was developed using python, and the algorithm consists in the correlation of proteolytic sites with their related secondary structure. We evaluated all nonredundant structures from human and viral proteins hosted in Protein Data Bank. Our initial results point to a tendency for the occurrence of cleavage in beta strands conformation, what would represent the presence of some structuration degree of the substrate before the degradation by the proteasome.

Finantial supported: Bill & Mellinda Gates Foundation, Fapergs, Capes.

[1] Núcleo de Bioinformática do Laboratório de Imunogenética, Departamento de Genética, Universidade Federal do Rio Grande do Sul, Postcode 91501-970, Brazil
{imunoinfo@gmail.com}

Differentially Expressed Genes During Salt Stress in Rice (*Oryza sativa* L.)

Daniel da Rosa Farias[1], Luis Willian Pacheco Arge[1], Marcelo Nogueira do Amaral[2], Rodrigo Danielowski[1], Letícia Carvalho Benitez[2], Eugenia Jacira Bolacel Braga[2], Antônio Costa de Oliveira[1] and Luciano Carlos da Maia[1]

Rice is the staple food of more than two thirds of the world´s population, it is the second most widely grown cereal in the world, and Brazil is the largest producer outside Asia. This cereal is sensitive to salinity, which affects one fifth of irrigated lands worldwide. Developments in molecular marker and bioinformatics analyses became important tools to understand the gene responses related to abiotic and biotic stresses. Thus, this study aims to identify differentially expressed genes (DEGs) present in QTLs associated to salt tolerance in rice (cv. "BRS Querencia"). RNA-seq technique was used to analyze the transcriptional profile and elucidate the DEGs of rice under salt stress, total RNA was extracted from leaves in stage V3, exposed to stress for 24 hours. The paired-end sequencing of the cDNA libraries was performed on a HiSeq 2500 R Illumina platform. For the analysis and visualization of the read qualities, the software FastQC was used. Thereafter, the program Trimmomatic was used to remove low-quality bases and adapters from each library. The reads were mapped against the rice reference genome from The Rice Annotation Project Database, using the software TopHat, which uses software Bowtie to mapping. Software HTseq was used to count reads, and identification of DEGs was performed using software R with edgeR packaged (Bioconductor repository). Expression levels were normalized by the method RPKM considering differentially expressed genes with $P<0.01$ by Fisher's exact test. After a literature search, nine QTLs associated with salt tolerance in rice were used in this study. To align the DEGs in QTLs regions, the software BLAST was used. A total of 40 DEGs identified by analysis of RNAseq associated with QTLs were found. Thereby, these genes may be important for marked-assisted selection in breeding programs, improving the precision and efficiency in the selection for salt tolerance in rice.

Research supported by the following funding agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Rio Grande do Sul (FAPERGS).

[1] Plant Genomics and Breeding Center, Universidade Federal de Pelotas, Caixa Postal, 354 fariasdr@gmail.com
[2] Department of Plant Physiology, Universidade Federal de Pelotas, UFPel.

Characterization of Transposable Elements in Leptopilina boulardi Genome

Filipe Zimmer Dezordi[1]
Alexandre Freitas da Silva[1]
Elgion Lucio da Silva Loreto[2]
Paulo Marcos Pinto[1]
Gabriel da Luz Wallau[3]

Transposable elements (TEs) are mobile genetic elements present in the genome of most organisms studied so far [1]. TEs are grouped into two types, Type I transposons which uses a DNA intermediate in their mobilization process, and Type II that uses an RNA intermediate [2]. The characterization of these elements in new sequenced genomes becomes important to better understand the genome evolution as the ratio between mobile and non-mobile genome components. This study aimed to characterize the mobilome present in the wasp Leptopilina boulardi genome. After obtaining the wasp genome sequences, bioinformatics analyzes were conducted with RepeatExplorer tool, which uses a new approach based on clusterization of highly repetitive reads from next generation sequencing platforms. Our preliminary data shows several TEs inhabit L. boulardi genome, both Type 1 and Type 2 TEs. After an initial characterization we analyzed the TEs clusters with CENSOR tools to determine which homologues are present in RepBase and we are going to use ORF Finder to search for open reading frames, Basic Local Alignment Search Tool to search similarity against the NCBI. We obtained a total of 53 clusters representing the TE's. Where the elements of the Type I represented a total of 2.823% of the genome while the Type II represented a total of 2.37%. The elements that were present in greater amounts were LTR-Gypsy, representing 1,921%. The abundance of each TEs family is highly variable in insect genomes but, in general, Type 2 elements are more abundant than Type 1 although large variations in the contribution of each TE order are observed [3], this parameter is not found in L. boulardi genome, where type 1 elements are found in greater number, suggesting that the abundance of TEs in this genome can be result of a different host genome evolutionary history than other insects.

We thank CAPES, CNPq and Fapergs.

[1]Laboratório de Proteômica Aplicada, UNIPAMPA
{Filipe Zimmer Dezordi, zimmer.filipe@gmail.com; Alexandre Freitas da Silva, alexfreitasbiotec@gmail.com; Paulo Marcos Pinto, paulomarcospinto@gmail.com}
[2]Departamento de Bioquimica e Biologia Molecular, UFSM
{Elgion Lucio da Silva Loreto, elgionl@gmail.com}
3 Departamento de Ecologia, Zoologia e Genética, UFPEL
{Gabriel da Luz Wallau, gabriel.wallau@gmail.com)

Potential use of GROMOS force field parameters for organic molecules

Marcelo Depólo Polêto[1]
Hugo Verli[1]

Techniques as docking and molecular dynamics have been applied around the world to rationally engineer molecules aiming to tackle different purposes. So far, force fields for organic compounds, like GAFF, have been designed to reproduce quantumly calculated parameters and, thus, do not take in account condensed phase terms. One recent exception was OPLS/AA applied to organic molecules (Caleman et al., 2011 - dx.doi.org/10.1021/ct200731v). With this in mind, the aim of this work is to create a small molecule force field based on GROMOS. Thus, charges, bonded and non-bonded parameters of GROMOS54A7 were used to create organic compounds topologies. After, we have simulated their properties using GROMACS 5.0.4, applying the same protocol as Caleman et al.. Therefore, density and enthalpy of vaporization were calculated for each compound, along with absolute error in relation to experimental values. Absolute error values were used to determine whether the quality of each topology was acceptable. When required, manual adjusts on topology were done to guarantee its confidence. So far benzen, 1H-pyrrole, fluorobenzen, 1,2-difluorobenzene, 1,3-difluorobenzen, 1,2,3,5-tetrafluorobenzen, pyridine and pyrimidine were succesfully parameterized. Regarding density, all compounds had absolute error below 5%, except 1H-pyrrole(5,05%) and pyrimidine (8.46%). Regarding enthalpy of vaporization, all compounds also had absolute error values below 5%, except 1H-pyrrole (9.30%), 1,2-difluorobenzen (5.76%) and 1,2,3,5-tetrafluorobenzen (10,38%). Moreover, other molecules are being built in order to expand our database and other properties will be included in our analyzes to guarantee a good description of more termodynamical properties. These preliminary results suggest the potential use of GROMOS43A7 parameters to create organic molecules with good agreement with their physical-chemical properties, and can lead to a rich database for drug design purposes.

Supported by: CNPq, CAPES and FAPERGS

[1] Centro de Biotecnologia, UFRGS, Caixa Postal 15005
{marcelodepolo@gmail.com; hverli@cbiot.ufrgs.br}

Genetic Mapping of Diseases with NGS using Big Data Analysis

Julio C. S. dos Anjos[1]
Junior F. Barros[1]
Raffael B. Schemmer[1]
Claudio F. R. Geyer[1]
Ursula Matte[2]

Abstract: The development of advanced genetic sequencing techniques has allowed creating what is called Next-Generation Sequencing (NGS) to provide new advances into genetics field. However, new problems related to processing of significant amounts of data were discovered and begin to turn out complexity to solve genetic sequencing problems. The necessity of new methods of processing grows the towards the NGS technology. Also a fast and accurate analysis is relevant in the context of diseases due to time for treatment. This study proposes the development of genetic notation systems through the MapReduce framework to allow disease analyses in an automated way to provide tools for researchers and doctors in a clinical environment.

Introduction

The information provided by DNA is crucial to improve the development of several areas of biological research. The studies about new pharmaceutical substances, foodstuffs, pesticides and agricultural products are clearly benefited from biotechnology and their insights [1]. One the most focused fields in terms of efforts is the clinical analysis, which is strongly and constantly improved by research, in case specially the detection of pathologies will be centered in this work. When looking a few years ago, DNA sequencing was considered an expensive technique, but it has changed since the Next-Generation Sequencing appears and a migration happened from Sanger. Next-Generation Sequencing or then called as NGS could provide a great solution looking the cost-effective relation in comparison to Sanger, although, NGS is a technique that produces a big amount of data after a sequencing process. As example shown in Table 1, NGS machines particularly those that use ion semiconductors (Ion Torrent PGM or Proton families) are able to sequence million of base pairs in few hours, computationally it means high throughput of data very quickly.

[1]Institute of Informatics, UFRGS, PoBox 15064
{jcsanjos@inf.ufrgs.br, jfbarros@inf.ufrgs.br, raffael.schemmer@inf.ufrgs.br, geyer@inf.ufrgs.br}
[2]Genetic Therapy Laboratory, Clinical Hospital of Porto Alegre, UFRGS
{umatte@hcpa.edu.br}

Table 1. Technical specifications of Ion Torrent sequencing machines.

	Machine Models			
Characteristics	PGM318	PI	PII	PIII
Sensor Number	~11 M	~165 M	~660 M	~1.2 B
Input Size	~2 GB	~10 GB	~32 GB	~64 GB
Execution Time	4~7 hrs	2~4 hrs	2~4 hrs	2~4 hrs
Average Read	400 BP	200 BP	100 BP	100 BP
Number of Reads	~5.5 M	~82 M	~330 M	~660 M

Key: M = Million B = Billion BP = Base Pairs GB = GigaBytes

Therefore, the techniques used nowadays for processing the unprecedent amount of data generated by NGS are limited and falls when lack the scalability or response to ease the diagnosis process. Several variations of human genome can be determined by massive parallel sequencing. However, many of these are not clinically relevant, the need of methods to discriminate between disease-causing mutation and normal genetic variability is a short run-time [4]. The purpose of this work is to present a model of system that can assist doctors in clinical diagnosis of patients by conducting an analysis of the genetic mutations contained in their DNA. The scientific goal is to provide an efficient and robust method for the genetic mapping of diseases through NGS Big Data.

Mutations, Polymorphism and Clinical Genetics

A mutation is defined as a change in the nucleotide sequence of an organism, it can be caused by an irreparable damage suffered by the genome, errors in the replication process or insertion/deletion of DNA fragments by mobile genetic elements. Several studies such as [5] suggests that mutation changes the proteins that a gene produces and it causes dangerous consequences for the organism. Polymorphism is a kind of mutation that takes place with a frequency greater than 1% in a population and can be divided into distinct, well defined classes. According [6] there are three classes of mutations, those alter the number of chromosomes in a cell (genomic mutations), those alter structures of specific chromosomes (chromosomal mutations) and mutations that change individual genes (gene mutation).

MapReduce

MapReduce is a programming framework that abstracts the complexity of parallel applications by partitioning and scattering data sets across hundreds or thousands of machines, and by bringing computation and data closer [2]. In case, the MapReduce is an attractive solution to imply as it offers a parallel distributed solution for processing in high volumes of data such as those produced by sequencers as it free certain difficulties from programmers when developing distributed systems. Beyond that, using Apache Hadoop implementation, it has its own solution for distributed file system called HDFS. Using MapReduce to develop an application applied to bioinformatics could provide a reduction of complexity [3] focusing efforts to solve problems.

Description of the Model

This study implements a genetic mapping by using MapReduce to analyze and process the amount of data generated by the sequencing process. Three stages compose this work through the Hadoop implementation. A filtering step executes a script to preprocessing that saves in memory the gene from the patient to form a single input. This filtering occurs before to begin the first phase of MapReduce processing. In the first stage, the patient's gene is sent to a distributed file system to be processed from nodes. In the second stage, each gene with its respective identifier is evaluated against to the reference genome to find a gene position that has some anomaly using CCDS database. In the second stage, if some anomaly was found, a key/value is emitted in memory for each node. Where the key is the patient identifier and the value, the gene position changed. This model assumes that the input file is formed by different kinds of genes and patients. The third stage is to compose the annotation by using the MapReduce processing. Figure 1 illustrates this process.

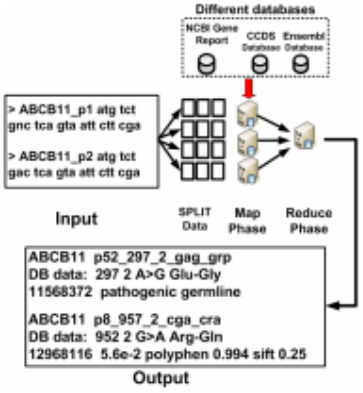


Figure 1. Flowchart of the proposal.

In the Map phase, a Combine execution compares the gene anomaly found against both Ensembl and Gene Report databases. The search process uses the position and the name of the gene to query the databases, if a pathology is reported, the Map emits a key/value pair intermediate. The key is formed by gene's position and the value within a tuple of a reference gene and patient's pathology. After, the reduce function emits a new key/value pair with the information associated about all pathologies found for each patient and saves on HDFS. Only mutations found in databases are written in the output, followed by messages from the associated pathological informations.

Conclusion

This study proposes a model for genetic mapping of diseases through Big Data analysis from different patients simultaneously in a scalable fashion, to support researchers and doctors in an automated way. MapReduce framework establishes a coherent model to the development of genetic notation solutions and provides a high level of parallelism in cluster or cloud environments.

References

[1] William J, T. and Palladino, M. A. (2012). Introduction to Biotechnology, volume 1. Pearson, 3rd edition.

[2] Dean, J. and Ghemawat, S. (2010). MapReduce - A Flexible Data Processing Tool. Communications of the ACM, 53(1):72–77.

[3] Zou, Q., Li, X.-B., Jiang, W.-R., Lin, Z.-Y., Li, G.-L., and Chen, K. (2014). Survey of MapReduce frame operation in bioinformatics. Briefings in Bioinformatics, 15(4):637–647.

[4] Frebourg, T. (2014). The challenge for the next generation of medical geneticists. Hum Mutat, 35(8):909–11.

[5] Johnsen, J. M., Nickerson, D. A., and Reiner, A. P. (2013). Massively parallel sequencing: the new frontier of hematologic genomics. Blood, 122(19):3268–3275. [6] Nussbaum, R., McInnes, R., and Willard, H. (2013). Thompson Genetics in Medicine. Elsevier Science Publishers B. V., 7th edition.

Analyzing Microarray Data Using Gene Regulatory Networks Cycles

Fabiane Cristine Dillenburg[1], Alfeu Zanotto-Filho[2], José Cláudio Fonseca Moreira[2], Leila Ribeiro[1], Luigi Carro[1]

Introduction

Gene expression provides information for building models of biological systems. Gene expression analysis comparing normal and neoplastic tissues have been used to identify genes associated with tumor genesis and potential therapeutic targets [1]. Genomic high-throughput technologies, such as microarrays, may considerably facilitate the molecular profiling of human tumors. Thousands of genes can now be analyzed using a single microarray hybridization chip [2]. The expression profile from a single tumor reflects the state of events of an individual malignancy at a certain time point. To generalize the findings and provide conclusive evidence for the involvement of a molecular alteration, it is often necessary to analyze several hundred tumors. Using traditional molecular pathology, such verification could take several months, or even years, to reach completion. To facilitate translational research in a large-scale manner, new techniques are needed. One of the main research areas in systems biology concerns the discovery of biological regulatory pathways or networks from microarray datasets [3]. A gene regulatory network (GRN) consists of a great number of genes whose expression levels affect each other in various ways. Computational models of GRNs can take a variety of forms, include models comprised of directed and undirected graphs. The process of constructing a regulatory network that explains some behavior of the cell using microarrays can be done in two steps: first, the set of genes that is thought to be relevant for this biological process is selected and measured in the microarray (for all the samples); then, the gene expression data is analyzed to generate graphs that represent the desired regulatory network. In this work, we present a new way of analyzing microarray datasets, based on the different kind of cycles found among genes of the GRN constructed using quantized data obtained from the microarrays. A cycle is a closed walk with all vertices (and hence all edges) distinct (except the first and last vertices). Thanks to the new way of finding relations among genes, a more robust interpretation of gene correlations is possible. Furthermore, the cycles help differentiate, measure and explain the phenomena identified in healthy tissue and diseased tissue. We use the proposed methodology to analyze the genes of three networks closely related with cancer - apoptosis, glucolysis and cell cycle - in tissues of the most aggressive type of brain tumor (Gliobastomamultiforme – GBM) and in healthy tissues. Because most patients with GBMs die in less than a year, and essentially no patient has long-term survival, these tumors have drawn significant attention [4].

Methodology

Firstly, we selected the main genes involved in the pathway of interest. The gene expression analysis comparing normal and GBM tissues was performed from previously published and characterized database comprising 276 GBM samples of all histology

1 Instituto de Informática, UFRGS, Porto Alegre, Rio Grande do Sul, Brazil. {fabiane.dillenburg, leila, carro@inf.ufrgs.br}

2 Departamento de Bioquímica - ICBS, UFRGS, Porto Alegre, Rio Grande do Sul, Brazil. {alfeuzanotto@hotmail.com, 00006866@ufrgs.br}

compared to 8 brain samples of non-neoplastic white matter tissue. The raw data for this study is available as experiment number GSE16011 in the Gene Expression Omnibus[2]. Experimental data used in the analysis are available in AffymetrixGeneChip Human Genome U133 Plus 2.0 Array format. The analyses of Affymetrix microarray data were performed using R[3] and Bioconductor[4]. Our analysis method consists of the following steps:

1) **Preprocessing of Affymetrix microarray data.** This step consists of importing the raw data in and summarizing the expression values per each probe set. The process of summarizing expression values is constituted of (a) background correction; (b) normalization; (c) summarization. All these operations are supported in the Bioconductor package *affy*.

2) **Annotation data.** The purpose of the annotation is to provide detailed information about the data. These operations are supported in the Bioconductor package *annotate* and *hgu133plus2.db*. We created a dataframe with the features names, the genes symbols and the summarized data samples. We then selected the data from the genes of interest through its symbol. These data were divided in GBM samples and control samples.

3) **Sigmoidal normalization.** This step reduces the influence of extreme values or outliers in the data without removing them from the dataset. Data are nonlinearly transformed by using a sigmoidal function [5] and the normalized values range from 0 to 1.

4) **Spearman's Correlation.** Correlation is used to discover sets of genes with similar expression profiles. Spearman's rank correlation coefficient is non-parametric and allows one to identify whether two variables (genes) relate in a monotonic function.

5) **Graphs.** The undirected graphs (representing the GRN) are constructed by computing a correlation coefficient for each pair of genes. If the coefficient is above a certain threshold and is statistically significant ($p<0.05$), the gene pair gets connected in the graph, if not, it remains unconnected. We used the R package *igraph* for manipulating undirected graphs. The adjacency matrix used is a matrix with correlation coefficients greater than the defined threshold.

6) **Cycles.** In order to seek the biological explanation of the observed gene association, we decided to look for cycles in the gene network. We used a C++ implementation of Johnson's algorithm [6] to find the cycles in the graphs. Feedback mechanisms are very common in biological networks. Our hypothesis is that negative feedback could allow one to find relations among genes that could help explain the stability of the regulatory process within the cell. Positive feedback cycles, on the other hand, can show the amount of imbalance a certain cell is suffering when in that state. The genes of interest are of two types: activators and inhibitors. We assume that a cycle of a graph is positive when the number of inhibitors in the cycle is zero. Similarly, it is said to be negative when the number of inhibitors in the cycle is greater than or equal to one.

Results

We use the proposed methodology to analyze the genes of three networks closely related with cancer - apoptosis, glucolysis and cell cycle. The results evidence differences between the GRNs of the three networks among the control samples and GBM samples. The cycles of the control graphs use about all the genes of each network; while the cycles of the GBM graphs using a small group of genes of each network. In apoptosis, only a few cycles

[2] http://www.ncbi.nlm.nih.gov/geo/
[3] http://www.r-project.org/
[4] http://www.bioconductor.org/

were found in the GBM graph, it would indicate that the cell cannot die [7]. In the cell cycle, the greater number of cycles has been found in GBM, which might indicate that the tumor has more active cell cycle mechanisms, since it is more proliferate [7]. Analyzing the most common genes found in the cycles, we observed that the t-test with 0.05 significance level indicated that there is no significant difference between the average of the gene expression level of the control samples and the GBM samples of some genes of the three pathways. Thus, there is a gain of information with the analysis using cycle to analysis of the gene expression level, whereas cycles highlight the difference between the control samples and the GBM samples.

Discussion

In literature, some articles focus on how to detect biologically meaningful modules [8] and recurring patterns called motifs [9] in networks. Our analysis is focused on the genes from a pathway so the goal was not to identify modules, pathways or motifs, but rather to better understand the relationships of the genes of the pathway of interest and its variations between samples of GBM and control to get insights on how alterations in the levels of inhibitors may affect the activation of the pathway based on target genes evaluation. The proposed approach innovates by using the existing cycles in the network for analysis, instead of using the connectivity of the whole network or the intramodular connectivity as the measure of node importance as other approaches do [10], thus providing a different and potentially fruitful strategy to analyze complex interactions in pathways.

References

[1] PARMIGIANI, G. et al. The Analysis Of Gene Expression Data: an overview of methods and software. In: PARMIGIANI, G. et al. (Ed.). The Analysis Of Gene Expression Data: methods and software. New York: Spring Verlag, 2003.

[2] STEKEL, D. Microarray Bioinformatics. Cambridge University Press, 2003.

[3] ALTAY, G.; EMMERT-STREIB, F. Inferring the conservative causal core of gene regulatory networks. BMC Systems Biology, [S.l.], v.4, p.132, 2010.

[4] MRUGALA, M. M. Advances and challenges in the treatment of GBM: a clinician's perspective. Discov. Med., [S.l.], v.15, n.83, p.221–230, 2013.

[5] PRIDDY, K. L.; KELLER, P. E. Artificial Neural Networks: an introduction. Bellingham: SPIE (The International Society for Optical Engineering) Press, 2005.

[6] JOHNSON, D. B. Finding All the Elementary Circuits of a Directed Graph. SIAM Journal of Computing, [S.l.], v.4, n.1, p.77–84, 1975.

[7] HANAHAN, D. WEINBERG, RA. Hallmarks of cancer: the next generation. Cell. v.144, n.5, p.646-74, 2011.

[8] LANGFELDER, P.; HORVATH, S. Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology, v.1, n.1, p.54, 2007.

[9] ALON, U. Network motifs: theory and experimental approaches. Nat Rev Genet, v.8, n.6, p.450-61, 2007.

[10] MA, S. et al. Incorporating gene co-expression network in identification of cancer prognosis markers. BMC Bioinformatics, v.11, n.271, 2010.

Molecular Dynamics of Stapled Peptides

Bianca Villavicencio[1]
Rodrigo Ligabue-Braun[1]
Hugo Verli[1]

Peptides are promising for drug development research due to their reduced size, their ability to traverse membranes, and a reasonable contact surface for interacting with target proteins. However, without its complete proteic framework, a peptideâs secondary structure can be easily destabilized, allowing the loss of its biologically relevant conformation. In α-helices, a strategy to address this issue is the addition of a molecular staple linking one or more turns of the helix through an all-hydrocarbon bridge. Studies of this type of structure yielded good results both in vitro and in vivo. Simultaneously, predictive models are still scarce, but would reduce research costs and allow a better understanding of stapled peptides and their potential use. One of such models is achieved by simulations by molecular dynamics, which renders information about conformational changes in the peptide. In order to allow their application in protein engineering, we parameterized and simulated both stapled and unstapled peptides, starting as either helices or coiled molecules with the potential to acquire defined structures. We evaluated staples of 6 or 8 carbons with different configurations (R-S). All simulations were carried out with the GROMOS54a7 force field in the Gromacs simulation suite. Results indicate that, while unstapled peptides tend to lose their helical form, stapled peptides tend to remain helical. In the case of initially coiled stapled peptides, the trend is to become helical. The type of staple used seems to influence the degree of helical content acquired. These alterations in secondary structure content seem to agree with experimental circular dichroism results. These results might allow the implementation of simulation protocols for stapled peptides, and also seem to describe with confidence the conformational behaviour of such peptides at the atomic scale. Further simulations with different force fields are being performed in order to obtain more data for comparison.

Supported by: CNPq, CAPES and FAPERGS

[1]Centro de Biotecnologia, UFRGS, Caixa Postal 15005
{bia.villavicencio@gmail.com; hverli@cbiot.ufrgs.br}

Entering Inflammaging pathways into Ontocancro 3.0

Laís Falcade[1]

Giovani R. Librelotto[1]

Rômulo Stringhini[1]

In the natural process of life, people are born, grow, age and die. The degeneration of the cells that occurs in this process is unconditionally, in fact, the inability to renew cell, and this failure is that live various diseases. Knowing that every human being older and, by consequence, are prone to more degrading inflammatory processes than in youth, the search for longevity is a fascination for many scholars. A topic that has been researched in recent years in this area is Inflammaging, which deals with the chronic inflammatory state that comes with aging. This work has promoted the development of an ontology oriented to the study of inflammaging being unified to Ontocancro. In the context of this work, ontology has the purpose to map the existing knowledge in the genetic framework of chronic inflammatory state elapsed by aging, such as Alzheimer's disease, Type 2 diabetes, COPD or pulmonary disorder Chronic Obstructive the Thyroid Cancer of Pancreas of Colon and Rectal Adrenocortical glands, as well as samples of each disease, metabolic pathways and genes present in this process. All data were defined based on current research literature. The dissertation presents results in a comparative analysis between the tracks of the natural process of cell development with inflammation pathways inserted after this study, and you can see that both are intertwined from the genetic interaction. To visually verify these connections used the String, a genetic link processor that provides graphs relationship between genes, showing the intensity ratio within a bonding means or more lanes. Thus, there was a comparative study of two routes; detected genes of intersection between them have intensive connections to a degree of 90% reliability.

[1]Programa de Pós-Graduação em Informática – Universidade Federal de Santa Maria

Alternative Splicing Analysis in Rice Under Iron Overload

Artur Teixeira de Araujo Junior[1]; Marcelo Nogueira do Amaral[2]; Luis Willian Pacheco Arge[2]; Daniel da Rosa Farias[2]; Railson Schreinert dos Santos[2]; Danyela de Cássia Oliveira[2]; Solange Ferreira da Silveira Silveira[2]; Eugenia Jacira Bolacel Braga[2]; Luciano Carlos da Maia[2]; Antonio Costa de Oliveira[3]

Introduction

Rice (Oryza sativa L.) is an important crop, being the second most produced cereal worldwide. In Brazil, rice is mainly produced in ecosystems called "terras baixas" (lowland). These conditions of waterlogging cause a large problem: iron overload. This problem culminates with large losses in crop production [1], possibly leading to a reduction of up to 100% in yield, depending on the concentration of reduced iron in soil solution and on the tolerance of rice cultivars [2]. Recent studies report that environmental stresses can affect the mechanism of alternative splicing, causing changes in the expression of different genes and affecting especially those involved in post-transcriptional modifications. Due the little information on this subject, this study aimed to evaluate and quantify the different junction sites and alternative splicing events in the transcriptome of rice cultivar BRS Querência under normal and high iron concentrations.

Materials and Methods

Seeds of rice (cv. BRS Querência) were germinated in a growth chamber for 7 days with a photoperiod of 16 hours and a temperature of 25 ± 2 °C. After this period, seedlings were transferred to plastic trays containing pre-washed sand and kept in a greenhouse with alternate irrigation (every 2 days) with nutrient solution [3] and water. Plants were subjected to iron overload at the seedling stage, adding 300 mg L^{-1} Fe^{+2} to the nutrient solution, for 24 hours. Untreated plants remained in normal nutrient solution for the same time, to be used as control. The total RNA was extracted from 100 mg of tissue (leaves) using of Plant RNA Reagent Purelink®. For the preparation of libraries was used TruSeq RNA Sample Preparation v2 (Illumina®) kit, following manufacturer's recommendations. RNA-Seq analysis was performed using HiSeq 2500 platform (Illumina®) with paired-end 2 x 100 reads (two reads of 100 bp). In order to evaluate read quality, the software FastQC Ver. 0.11.2 was used. After that, the software Trimmomatic Ver. 0.32 [4] was used toremove the adapters and low quality base reads of each libray. Then the reads were mapped using O. sativa Nipponbare (IRGSP build 4.0) as reference in TopHat Ver. 2.0.11 [5]. After alignment, the software MapSplice [6] and SpliceGrapher [7] were used to analyze of the junction sites and alternative splicing events, respectively. Predicted Rice Interactome Network - PRIN [8] were used to demonstrate probable protein-protein interactions.

[1] Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900 (arturtaj@hotmail.com)

[2] Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900

[3]Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900 (acostol@cgfufpel.org)

Results and Discussion

A total of 127,781 alternative splicing junctions sites were identified in plants under normal conditions (control), while 123,682 were found in stressed plants. When these results are compared with the literature, a significant difference in the proportion of these junction sites into the three main categories described is found (Figure 1A). These results indicate that there is no big change in splicing processes due to iron overload stress. As expected most common sites are shown to be the canonical, involving 98,85% and 98,91% in control and stress condition respectively, consistently with other results in the literature [9, 10]. Also that we have the same-canonical (0,73% in control and 0,70% in stressed) and non-canonical junctions sites (0,42% in control and 0,40% in stressed). A deeper alternative splicing analysis showed that the control plants presented a higher number of events (23,307) when compared to the treated ones (8,244). Furthermore it is interesting to note that intron retention was the most frequent event in both situations, with 10,281 (44.1%) for control and 3,909 (47.4%) for stressed (Figure 1B and C). Similar results were also detected in others studies, not just for rice but also for other plants species [10, 11]. The second major event is the alternative 3' splice site (5,261/22.6% and 1,802/21.9%), followed by exon skipping (4,413/18.9% and 1,429/17.3%) and the alternative 5' splice site (3,352/14.4% and 1,104/13.4%). Some studies demonstrate that the intron retention can even form truncated proteins and these proteins can make important functions in the adaptation of plants to stress conditions [12].

The alternative splicing influences are present in almost all aspects of protein functions, making it a central element of gene expression regulation. In this case, changes promoted by alternative splicing in the expression of proteins related to post-translational modifications were analyzed. These proteins are responsible for diversifying the functions of others proteins and for the dynamical coordination of signaling networks [13]. Here we highlight that, 25 differentially expressed proteins were found related to post-translational modifications, and from these, 20 were downregulated, while 5 were upregulated. When these proteins were analyzed, kinases and phosphatase were found to be the most common classes, where 14 assignments were found. These proteins are responsible for two important biological processes: phosphorylation with 14.29% assignments (3 downregulated) and dephosphorylation with 85.71% assignments (14 downregulated and 4 upregulated), and the most common molecular functions that are transferase activity and ATP binding both with 16 downregulated and 4 upregulated assignments. PRIN presented data about interactions with other differentially expressed proteins for just one locus (LOC_Os02g34600) in this experiment. Results demonstrated that alterations in the expression of this protein correlates with other 6 (Figure 1D), having 5 negative correlations with co-expression between -0.1879 and -0.021, and 1 positive correlation (0.1756). As demonstrated in Figure 1E the metabolomic correlation network, indicates protein-folding metabolism which presented negative co-expression with the post-translational modifications, and post-translational modifications metabolism, presenting positive co-expression with the RNA processing and negative co-expression with ribosomal protein synthesis and another non-assigned metabolism.

From these results one can conclude that BRS Querência demonstrated change in alternative splicing and the post-translational modification proteins promoted by the

iron stress. This initial data is important for the understanding on how plant performs their responses to stress and how this response may be related to the t olerance of this cultivar. The identification of transcripts can aid its use in genetic engineering and marker development.

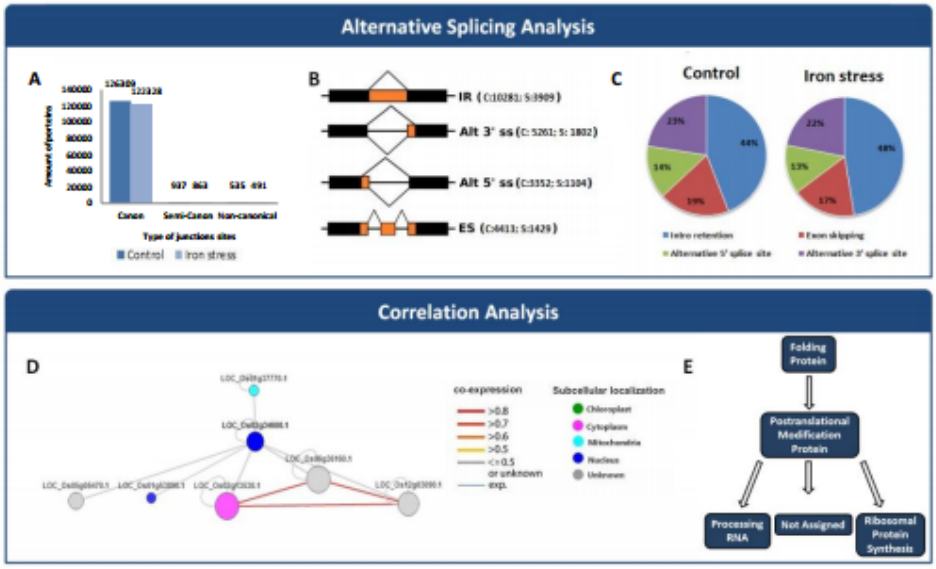


Figure 1 Alternative splicing analysis. (A) Types of alternative splicing junctions sites.(B) Schematic representation of the most frequent splicing event identified in the v2 prediction: intron retention (IR), alternative 3' splicing site (Alt 3' ss), alternative 5' splicing site (Alt 5'), exon skipping (ES). The number of events is reported between brackets in control (C) and stress condition (S). (C) Pie chart showing the percentage distribution of alternative splicing events. (D) Protein-protein interactions network. (E) Correlation metabolomics network.

References

[1] VAHL, L.C., Iron toxicity in rice genotypes irrigated by flooding. Ph.D. Thesis. Federal University of Rio grande do Sul, Porto Alegre,1991.

[2] SAHRAWAT, K. L. Iron toxicity in wetland rice and the role of other nutrients. Journal of Plant Nutrition, v. 27, p. 1471-1504, 2004.

[3] YOSHIDA, S.; FORNO, D.A.; COCK, J.H.; GOMEZ, K.A. Laboratory manual for physiological studies of rice. The Philippines: International Rice Research Institute(IRRI), 1972. 3v.

[4] BOLGER, A.M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, Reino Unido, v.30, n.15, p.2114 -2120, 2014.

[5] TRAPNELL, C.; PACHTER, L.; SALZBERG, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, Reino Unido, v. 25, n. 9, p. 1105-1111, 2009.

[6] WANG, K.; SINGH, D.; ZENG, Z.; COLEMAN, S.J.; HUANG, Y.; SAVICH, G.L.; HE,X.; MIECZKOWSKI, P.; GRIMM, S.A.; PEROU, C.M; MACLEOD, J.N.; CHIANG, D.Y.;

PRINS, J.F.; LIU, J. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research, Reino Unido, v.38, n.18, p.1-14, 2010.

[7] ROGERS, M.F.; THOMAS, J.; REDDY, A.S.N.; BEN-HUR, A. SpliceGrapher: Detecting patterns of alternative splicing from RNA-seq data in the context of gene models and EST data. Genome Biology, Reino Unido, v. 13, n.4, p.1-17, 2012.

[8] GU H.B., ZHU P.C.; CHEN M. PRIN: a pedicted rice interactome network. BMC Bioinformatics, 2011.

[9] AMARAL, M.N.; ARGE, L.W.P; BENITEZ, L.C.; MORAES, G.P.; MAIA, L.C.; BRAGA, E.J.B. Eventos de splicing alternativo associados ao estresse salino em plantas de arroz. ENPOS, 2014.

[10] VITULO, N.; FORCATO, C.; CARPINELLI, E.C.; TELATIN, A.; CAMPAGNA, D; D'ANGELO, M.; ZIMBELLO, R.; CORSO, M.; VANNOZZI, A.; BONGHI, C.; LUCCHIN, M; VALLE, G. A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biology, Reino Unido, v.14, n.99, p.1-16, 2014.

[11] LU, T., LU, G., FAN, D., ZHU, C., LI, W., ZHAO, Q., FENG, Q., ZHAO, Y., GUO, Y., LI, W. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res. 20, 1238–1249, 2010.

[12] MATSUKURA, S.; MIZOI, J.; YOSHIDA, T.; TODAKA, D.; ITO, Y.; MARUYAMA, K.; SHINOZAKI K.; YAMAGUCHI-SHINOZAKI K. Comprehensive analysis of rice DREB2 type genes that encode transcription factors involved in the expression of abiotic stress-responsive genes. Molecular Genetics and Genomics, Alemanha, v.283, n.2, p.185-196, 2010.

[13] WANG, Y.C.; PETERSON, S.E.; LORING. J.F. Protein post-translational modifications and regulation of pluripotency in human stem cells. Cell Research, 2014, 24:143-160.

Workflow for Illumina Amplicon Analysis of Soil Communities under Pollutant Stress

Daiana Lima-Morales[1,2], Ruy Jáuregui[1,3], Amélia Camarinha-Silva[1,4], Ramiro Vilchez-Vargas[1,5], Dietmar H. Pieper[1]

There is still a general lack of knowledge how microbial communities react under pollutant stress, further more it is known that microorganisms enriched in the laboratory often don't play an important role in in-situ biodegradation. Disregarding the need of cultivation, molecular techniques allow studying the microbial diversity in-situ. To analyze which microorganisms are important in pollutant soils, we upgraded a high throughput method using the Illumina platform, to amplify the V5/V6 region of the 16S rRNA gene. In total 2.003.786 reads were obtained. Only reads of a minimum of 115 nt in length (29 nt of primer/barcode and 86 nt of 16S rRNA gene sequence) were considered. Truncated reads that had an N character, mismatches within primers and barcodes or more than 8 homopolymer stretches were discarded. All sequences, totaling 1.671.472 reads, were split into 63 files according to their unique barcode and collapsed into unique representative reads. Using Mothur, these reads were pre-clustered into 24,000 unique representative reads and afterwards filtered. A representative reads was kept if: a) it was present in at least one sample in a relative abundance >0,1% of the total sequences of that sample or b) it was present in at least 3 samples or c) it was present in a copy number of at least 10 in at least one sample. Phylotypes were then generated by clustering at 98% similarity using USEARCH and assigned to a taxonomic affiliation based on RDP classifier . Applying this workflow we could reduce the number of representative reads to 501 phylotypes, a computational manageable number, without compromising the fine scale soil community composition. Out of 501 phylotypes, 470 could be identified to phylum level, 432 to class level, 361 to order level and 312 to family level.

Financial Support: CNPq-Conselho Nacional de Desenvolvimento Científico e Tecnológico-Brazil and BACSIN from the European Commission.

[1] Microbial Interactions and Processes Research Group (MINP), Department of Medical Microbiology, Helmholtz Centre for Infection Research, Braunschweig, Germany

Current addresses: Laboratório de Pesquisa em Resistência Bacteriana (LABRESIS), Centro de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, Brazil; AgResearch, Palmerston North, New Zealand; Institute of Animal Nutrition, University of Hohenheim, Stuttgart, Germany; Laboratory of Microbial Ecology and Technology (LabMET), University of Ghent, Ghent, Belgium

ExOmin: Boosting Variant Priorization Identified by Whole-Exome Sequencing

Pedro B. Marin[1], Delva Leão[2], Junior F. Barros[1], Julio C. S. Anjos[1], Filippo P. Vairo[2], Jonas A. M. Saute[2], Claudio R. Geyer[1], Ursula Matte[2]

The Whole-Exome Sequencing (WES) is one of the main applications of Next-Generation Sequencing increasingly used into medicine to elucidate the molecular basis of mendelian diseases without defined etiology by tradicional methodologies. The strategy to prioritize the causal genetic variant among thousands found on a WES analysis as well as the connection between phenotype and genotype specifically responsible by the condition investigated are crucial and laborious tasks. Such complexity has stimulated the developing of methodologies to assist this process. The majority of available prioritization tools are paid, overly complex or do not meet all aspects needed to perform an integrated analysis. Seeking to attend a real demand of medical geneticists of Clinical Hospital of Porto Alegre (HCPA), that face in prioritization step a relevant obstacle to define the diagnosis of patients attended at Genetic Medical Service (SGM), we propose the developing of ExOmin tool. This tool will comprise different modules and will use a strategy designed to combine variant annotation, variant prioritization, phenotype/genotype relation and selection of causal variant among probable variants. The main modules will be: 1)annotation of .VCF file with the open-source software ANNOVAR; 2) prioritization of variants using heuristic ilters; 3) integration of phenotype/genotype through OMIM database; 4) sorting of probable variants using conservation and pathogenicity predictors. ExOmim will be implemented as a workflow, where a .VCF file processed in different modules will result in fewer candidates variants to be defined as responsible for observed clinic features. It's expected that these modules can be used independently as required by the investigator. Once implemented, we hope that ExOmin boost the process of WES analysis at SGM/HCPA and that patients without defined diagnosis attended by this facility have their diagnoses with greater accuracy and in less time.

[1] Instituto de Informática, UFRGS, Caixa Postal 15064
{Pedro B. Marin, pedrobmarin@gmail.com; Junior F. Barros, jfbarros@inf.ufrgs.br;
Julio C. S. Anjos, julio.c.s.anjos@gmail.com; Claudio R. Geyer, geyer@inf.ufrgs.br}
[2] Serviço de Genética Médica, Hospital de Clínicas de Porto Alegre, Caixa Postal 15039
{Delva Leão, dleao@hcpa.edu.br; Fillipo P. Vairo, fvairo@hcpa.edu.br; Jonas A. M. Saute,
jsaute@hcpa.edu.br; Ursula Matte, umatte@hcpa.edu.br}

Bioinformatics applied to QTLs related to absorption and accumulation of arsenic in rice

Railson Schreinert dos Santos[1]
Eduardo Venske[1]
Artur Teixeira de Araújo Júnior[1]
Daniel da Rosa Farias[1]
Antônio Costa de Oliveira[1]

Introduction

Rice (Oryza sativa L.) is a staple food crop for half the world's population. This cereal is an important source of carbohydrates and other nutrients to the population and ha s a high socioeconomic importance to the producing countries. Due to its high consumption, the presence of toxic elements such as arsenic in its grains can be of concern. Studies have shown that rice is a major source of arsenic to humans through diet [1, 2], with a range of genotoxic effects [3, 4]. Countries such as Europe, United States [5, 6] and even Brazil [7], can have dangerous arsenic levels in rice grains. Arsenic is considered the most toxic xenobiotic and a carcinogen of the highest class of non-metals. Therefore, there is no minimum content that may be considered safe or harmless [8]. Breeding efforts to lower arsenic content in rice grains seem to be promising, since different studies show significant genetic variability for both arsenic accumulation and speciation. QTL (Quantitative Trait Locus), physiological processes, and genes responsible for these answers have also been studied [3, 9, 10]. Here we highlight one important QTL related to the change in the concentration of arsenic in shoots of seedlings. This QTL, generically called AsS, is flanked by RM318 and RM450 markers on rice chromosome 2, and accounts for 24.4% of the expected phenotypic variation [9]. The present investigation aimed to evaluate the use of some bioinformatics tools to analyze AsS, with the purpose of identifying genes related to the change in the concentration of arsenic in rice seedlings.

Materials and Methods

In order to identify the genes within the AsS locus (chr2:28634200..29637600) the genomic sequences of the MSU Rice Genome Annotation Project [11] has been used. The 145 identified genes were subjected to a differential analysis in Genevestigator [12] using the database obtained from a study of transcriptional expression in roots of cultivars with contrasting phenotypes for tolerance to arsenate [13].

[1]Plant Genomics and Breeding Center, UFPEL, PO Box 354

{railsons.faem@ufpel.edu.br; eduardo.venske@yahoo.com.br; arturtaj@hotmail.com; fariasdr@gmail.com; acostol@cgfufpel.org}

The changes of expression were evaluated using a p value ≤ 0.05 to select genes differentially expressed when under stress. The genes that showed to change its expression under stress had their similarity of expression evaluated using hierarchical clustering in the same experiment [13]. These genes were then evaluated for transcriptional expression in different organs, using the databases of Kudo et al., 2013 [14] and Norton et al., 2008 [15].

Results and Discussion

Analysis of the chromosomal region of AsS demonstrated the presence of 145 genes in this locus (data not shown). Genes that had transcriptional expression data available in the experiment of Norton et al., 2008 [13] were evaluated, and 3 loci were considered differentially expressed when plants were under 1 ppm of arsenate. The results of differential expression analysis are shown in Figure 1.

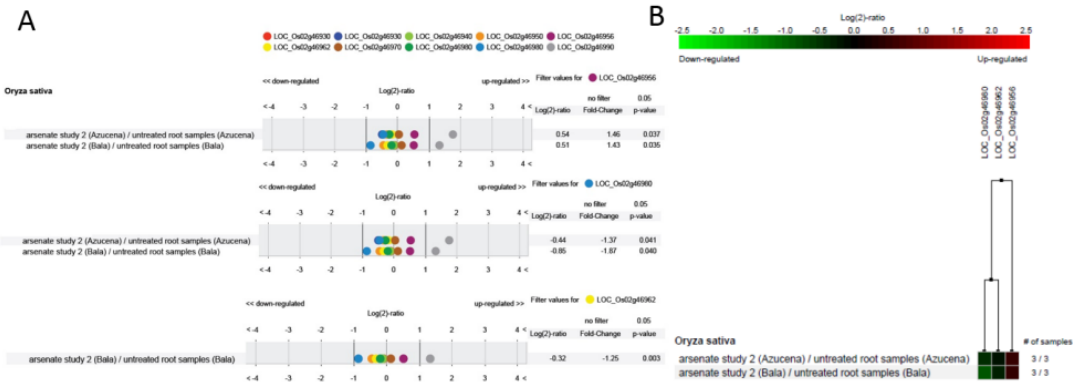


Figure 1. Differential expression of genes at the locus AsS. (A) Differential expression in plants under arsenate stress (1 ppm); (B) Hierarchical clustering analysis.

An increase in expression of LOC_Os02g46956 (chr2:28669044..28672392) and repression of LOC_Os02g46980 (chr2:28683849..28681681) and LOC_Os02g46962 (chr2:28673002..28676872) were observed. It is important to highlight that the last locus was only repressed in Bala, a tolerant genotype. To get an idea of the expression of these genes in different organs another analysis has been performed, generating the graph of Figure 2.

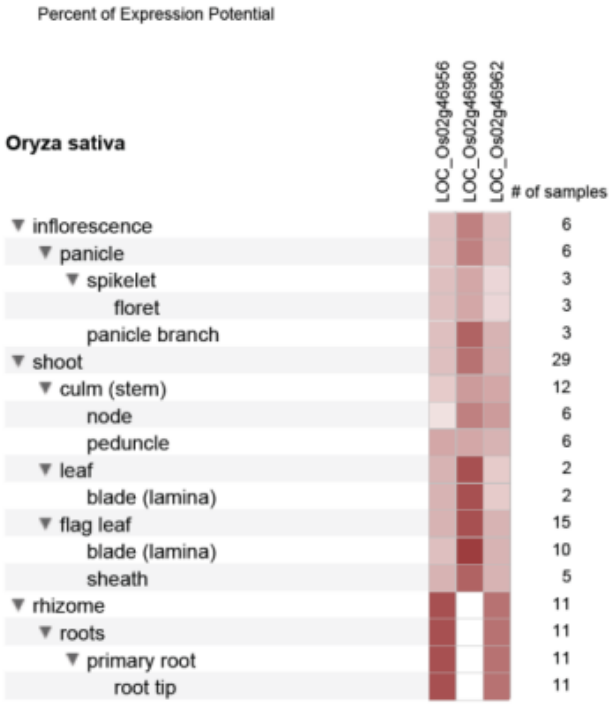


Figure 2. Expression of the three differentially expressed genes along rice anatomy.

The analysis of Figure 2 shows high expression potential of LOC_Os02g46956 and LOC_Os02g46962, another indicative of the probable importance in response to arsenate stress. Further studies involving characterization of these genes through structure prediction and modifications of these with molecular biology tools are yet to be conducted. The search for mutants and characterization these, are other important steps in elucidating the role of these genes and aid plant breeding. The results presented here show great utility of bioinformatics in the analysis of QTL related to absorption and accumulation of arsenic in rice plants. Results obtained from the analysis this and of other QTLs, also carried out by our laboratory (unpublished data), should help the understanding of the dynamics of arsenic in rice plants and their breeding to obtain safer crops and reducing the risk of cancer due to the intake of arsenic in food.

References

[1] Heinkens A. Arsenic Contamination of Irrigation Water, Soil and Crops in Bangladesh:

Risk Implications for Sustainable Agriculture and Food Safety in Asia. Bangkok: FAO, 2006.

[2] Meharg AA, Williams PN, Adomako E, Lawgali YY, Deacon C, Villada A, et al.

Geographical variation in total and inorganic arsenic content of polished (white) rice.

Environmental science & technology, vol. 43, no. 5, pp. 1612-7. 2009.

[3] Meharg AA and Zhao FJ. Arsenic and rice. Netherlands: Springer Netherlands; vol. 172, pp. 2012.

[4] Banerjee M, Banerjee N, Bhattacharjee P, Mondal D, Lythgoe PR, Martínez M, et al. High arsenic in rice is associated with elevated genotoxic effects in humans. Sci Rep, vol. 3, 2013.

[5] Williams PN, Price AH, Raab A, Hossain SA, Feldmann J and Meharg AA. Variation in arsenic speciation and concentration in paddy rice related to dietary exposure. Environmental science & technology, vol. 39, no. 15, pp. 5531-40, 2005.

[6] Zavala YJ, Duxbury JM. Arsenic in rice: I. Estimating normal levels of total arsenic in rice grain. Environmental science & technology, vol. 42, no. 10, pp. 3856-60, 2008.

[7] Batista BL, Souza JM, De Souza SS and Barbosa F, Jr. Speciation of arsenic in rice and estimation of daily intake of different arsenic species by Brazilians through rice consumption. Journal of hazardous materials, vol. 191, n. 1-3, pp. 342-8. 2011.

[8] Smith AH, Lopipero PA, Bates MN and Steinmaus CM. Public health. Arsenic epidemiology and drinking water standards. Science, 296, no. 5576, pp. 2145-6. 2002.

[9] Zhang J, Zhu YG, Zeng DL, Cheng WD, Qian Q and Duan GL. Mapping quantitative trait loci associated with arsenic accumulation in rice (Oryza sativa). The New phytologist, vol. 177, no. 2, pp. 350-5. 2008.

[10] Norton GJ, Pinson SR, Alexander J, McKay S, Hansen H, Duan GL, et al. Variation in grain arsenic assessed in a diverse panel of rice (Oryza sativa) grown in multiple sites. The New phytologist, vol. 193, no. 3, pp. 650-64. 2012.

[11] Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Research,vol. 35, pp. D883-D887, 2007.

[12] Zimmermann P, Hennig L and Gruissem W. Gene-expression analysis and network discovery using Genevestigator. Trends in plant science. vol. 10, n. 9, p. 407-9. 2005.

[13] Norton GJ, Nigar M, Williams PN, Dasgupta T, Meharg AA, Price AH. Rice–arsenate interactions in hydroponics: a three-gene model for tolerance. Journal of Experimental Botany, vol. 59, no. 8, pp. 2277-2284. 2008.

[14] Kudo T, Akiyama K, Kojima M, Makita N, Sakurai T and Sakakibara H. UniVIO: a multiple omics database with hormonome and transcriptome data from rice. Plant & cell physiology, vol. 54, no. 2, pp. e9. 2013.

[15] Norton GJ, Aitkenhead MJ, Khowaja FS, Whalley WR and Price AH. A bioinformatic and transcriptomic approach to identifying positional candidate genes without finemapping: an example using rice root-growth QTLs. Genomics, vol. 92, no. 5, pp. 344-52.

2008.

Functional analysis of microRNA and transcription factor co-regulatory network in pathological cardiac hypertrophy

Mariana R. Mendoza[1], Adriano V. Werhli[2], Graziela H. Pinto[1,3],Daiane Silvello[1,3], Carolina R. Cohen[1,4], Nadine O. Clausell[1,3], Luis E. P. Rohde[1,3], Andréia Biolo[1,3]

Introduction

Pathological cardiac hypertrophy is the increase in heart mass that occurs as response to pressure or volume overload in disease settings, or to myocardial infarction and cardiomyopathies. Although compensatory at first stage, sustained cardiac hypertrophy is detrimental and increases the risk of heart failure. Several well characterized structural, functional and molecular changes underlie this process, including cell death, fibrosis, cardiac dysfunction and altered gene expression. Recently, microRNAs (miRNAs) emerged as important post-transcriptional regulators of gene expression and have been implicated in cardiac development and several cardiovascular disorders [3]. Although their involvement with cardiac hypertrophy has been discussed in literature [2], their specific regulatory mechanisms, as well as the patterns of their cooperation in the synergistic regulatory network with transcription factors (TFs) have rarely been studied in this scenario. This study aims at constructing and analyzing the miRNA-TF co-regulatory network involved in cardiac hypertrophy using a bioinformatics approach, and help decipher major regulators contributing to this phenotype and their respective regulatory mechanisms and biological functions.

mRNA and miRNA expression profiles

Large-scale expression profiling of mRNAs and miRNAs related to an in vitro model of cardiac hypertrophy were downloaded from Gene Expression Omnibus (GEO) database (accession numbers GSE60291 and GSE60292). The study was carried out to identify RNAs in human cardiomyocytes that show differential expression upon induction of hypertrophy, assessing expression levels of miRNAs and mRNAs in triplicates for controls and endothelin 1 (ET-1) stimulated human cardiomyocytes [1].

[1]Laboratório de Pesquisa Cardiovascular, Centro de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre, Porto Alegre, RS, Brazil. E-mail address of corresponding author (MRM): {mrmendoza@inf.ufrgs.br}
[2]Centro de Ciências Computacionais, Universidade Federal do Rio Grande (FURG), Rio Grande, RS, Brazil
[3]PPG em Ciências da Saúde: Cardiologia e Ciências Cardiovasculares, Universidade Federal do Rio Grande do Sul,Porto Alegre, RS, Brazil
[4]PPG em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

Differential expression and functional enrichment analysis

To collect genes and miRNAs involved in the physiopathology of cardiac hypertrophy, we performed differential expression analysis based on the large-scale expression profiles from GEO. Original mRNA expression profile (GSE60291) was analyzed with affy and limma R packages, adopting background correction and quantile normalization by Robust Multi-array Average algorithm. Genes were considered differentially expressed between both groups if presenting absolute log2 fold changes greater than 1.5 and a false discovery rate (FDR) < 0.05. Differential expression analysis of miRNAs (GSE60292) was carried out with DEQSeq R package. MicroRNAs with minimum up- or down-regulation ratio of 1.5 (0.58 in a log scale) and FDR < 0.05 were considered significantly different between both groups. Following this analysis, we performed functional enrichment of differentially expressed genes (DEGs) using Gene Set Enrichment Analysis (GSEA) to extract biological insights such as the main molecular mechanisms deregulated upon induction of hypertrophy.

Construction and analysis of miRNA-TF co-regulatory network

A miRNA-TF co-regulatory network was constructed using a compilation of experimentally verified and confidentially predicted miRNA-gene (including TF genes), TF-miRNA and TF-gene interactions. Interactome data were retrieved from the following sources: i) ChipBase, HTRIdb and ORegAnno for validated TF-gene interactions; ii) miRTarBase and starBase for validated miRNA-gene interactions; iii) TargetScan for predicted miRNA-gene interactions; iv) ChipBase for validated TF-miRNA interactions. Only interactions in which both regulator and target were found to be differentially expressed in cardiac hypertrophy were considered for network construction. Once the network was generated, its structural properties, such as node degree and betweenness, were analyzed to investigate central genes contributing to network stability and communication. We defined hub nodes as the top 5% highest-degree nodes and bottleneck nodes as nodes with betweenness value standing one standard deviation above the mean. Finally, we analyzed the patterns of the interplay between miRNAs and TFs in the synergistic regulatory network by identifying significant network motifs using the FANMOD tool [4].

Differentially expressed genes and miRNAs

Based on GEO expression profiling datasets, we identified 496 differentially expressed genes (DEGs), including 24 TFs, and 273 differentially expressed miRNAs (DEMs) between both groups. A total of 201 genes, 16 TFs and 143 miRNAs were up-regulated upon induction of cardiac hypertrophy, while a total of 273 genes, 6 TFs and 130 miRNAs were down-regulated. Among the most deregulated genes and miRNAS, we identified significant overexpression of NPPB, ACTA1, CCL4, SERPINE1, SOX9, miR-1322, miR-21, miR-221 and miR-208a. According to GSEA, up-regulated genes are significantly enriched in regulation of MAPK activity, inflammatory and immune response, cytokine production and dephosphorylation, while down-regulated genes are mainly involved in distinct cell cycle phases,

including DNA replication, mitosis and sister chromatid segregation.

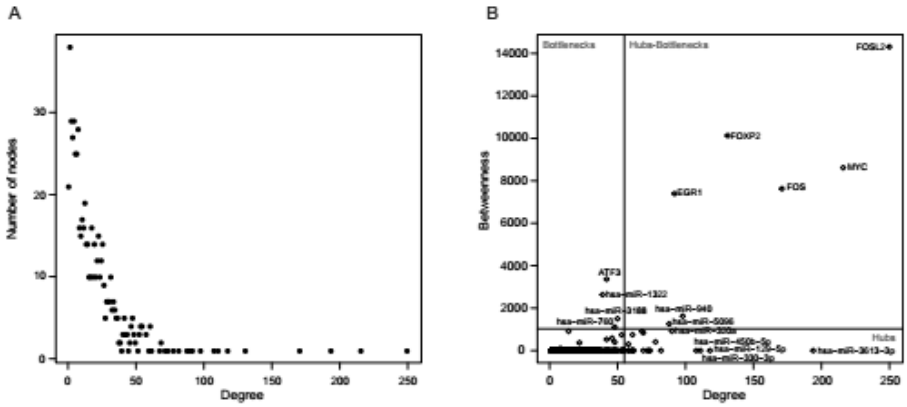


Figure 1. A) Degree distribution of the constructed miRNA-TF co-regulatory network. B) Degree vs. betweenness plot, defining hubs, bottlenecks and hub-bottlenecks nodes.

The cardiac hypertrophy miRNA-TF co-regulatory network

The miRNA-TF co-regulatory network included 6353 interactions among 617 nodes, recruiting 415 genes, 24 TFs and 178 miRNAs from those differentially expressed. Results of the node degree distribution reflected a scale-free topology, in which most nodes had low degrees and only a few nodes had high degrees (Fig. 1-A). Therefore, hub nodes might play major roles in network stability. Analysis of centrality measures identified several high degree nodes that may represent important network regulators, such as TFs MYC, FOSL2, FOS and EGR1, and miRNAs miR-3613-3p, miR-129-5p, miR-320a and miR-330-3p. Among the central nodes, we found that several hubs are also network bottlenecks, thus playing essential role in the control of information flow (Fig. 1-B). Network motif analysis identified significantly overrepresented patterns of interactions between TFs and miRNAs in the co-regulatory network (Fig. 2). A total of 18 motif types with size 3 were found englobing miRNA-TF synergistic regulation, among which we identified miRNA-mediated feedforward loop (FFL) regulation (Fig. 2-A), miRNA-mediated cascade regulation (B), TF-mediated FFL regulation (C), co-regulation (D) and miRNA simultaneous regulation (E). P-values were computed based on 1000 random simulations, comparing the frequency of motifs between randomly generated networks and the constructed network. Further analysis of these motifs and their constituents are needed to investigate common associations between miRNA and TF regulation.

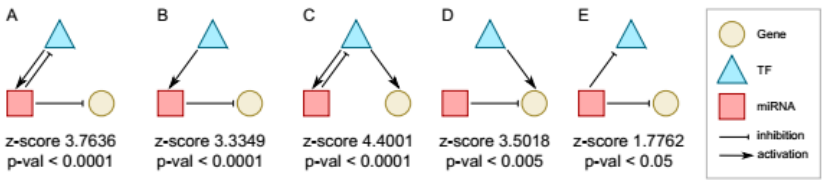


Figure 2. Over-represented motifs in the constructed miRNA-TF co-regulatory network.

Discussion

In this work we constructed a miRNA-TF co-regulatory network involved in the development of pathological cardiac hypertrophy, which allows us to analyse major regulators and regulatory patterns contributing to this mechanism and better understand its physiopathology. We identified important network regulators, such as TFs MYC, FOSL2, EGR1 and miRNAs miR-3613-3p, miR-129-5p, miR-330-3p, as well as significant miRNA-TF synergistic regulatory motifs that may be relevant for a better understanding of the molecular changes underlying cardiac hypertrophy. Further analysis of these data and a deeper investigation of network motifs and structural properties may help reveal common and essential patterns of interplay between miRNAs and TFs, as well as elucidate the regulatory role of miRNAs in pathological cardiac hypertrophy.

References

[1] P. Aggarwal, A. Turner, A. Matter, and et al. RNA expression profiling of human iPSCderived cardiomyocytes in a cardiac hypertrophy model. PLoS ONE, 9(9):e108051, 09 2014.

[2] P. A. Da Costa Martins and L. J. De Windt. MicroRNAs in control of cardiac hypertrophy. Cardiovascular Res., 93(4):563–572, 2012.

[3] V. Divakaran and D. L. Mann. The emerging role of microRNAs in cardiac remodeling and heart failure. Circ Res., 103(10):1072–1083, 2008.

[4] S. Wernicke and F. Rasche. Fanmod: a tool for fast network motif detection. Bioinformatics, 22(9):1152–1153, 2006.

In Silico Analysis of Gene Expression Profiling in Genes Linked to Autophagy and Pluripotency in Microarray Databases of IPSC, ESCS and Somatic Cells

PEREIRA, MB[1]
DALBERTO, TP[1]
LENZ, G[1]

Induced pluripotent stem cells (iPSC) have been widely explored as therapeutic tools in regenerative medicine and drug screening. The efficiency of the cellular reprogramming technique for the iPSC generation and of the cell differentiation process seem to be related to changes in energetic metabolism of these cells. Autophagy, in turn, appears to be important to the success of reprogramming. The identification of specific metabolic pathways and factors that regulate the destination of stem cells is important to the efficiency of reprogramming and to control the differentiation and destination of iPSCs. In this study, the in silico analysis of the correlation between groups of pluripotency genes and autophagy genes was performed using 12 microarray databases including ESCs, iPSCs and somatic cells. Samples from each database were grouped into clusters and the analysis of the correlation between groups of pluripotency genes and autophagy genes was done for the different cell types. Our findings indicate that pluripotency genes, such as Sox2, c-Myc and Lin28, are correlated with important genes involved in the autophagy induction in ESCs and iPSCs, suggesting the interaction between these two pathways and contributing to the better understanding of the relationship between them and their respective signaling pathways.

[1]Laboratório de Sinalização e Plasticidade Celular, Instituto de Biociências, UFRGS, Caixa Postal 9999
{Mariana Brutschin Pereira, mbrutschin@gmail.com} {Tiago Pires Dalberto, tiagodalberto@hotmail.com} {Guido Lenz, gulenz@gmail.com}

Multi-objective optimization in Bioinformatics

M. Villalobos Cid[1]

M. Inostroza-Ponta[1]

Many problems in bioinformatics and computational biology can be formulated as optimization problems: motifs discovery, expression data analysis, sequence alignment problems, gene regulation networks, protein structure prediction and recognition, single nucleotide polymorphism problem (SNP), drugs design, directed evolution analysis and optimization of biochemical processes (Biochemical informatics), among others. In general, single objective optimization problems are solved using algorithms based on techniques like linear and nonlinear programming, mixed integer programing, bilevel and dynamic optimization. In bioinformatics, most of the optimization problems belong to the NP-hard class of problems, i.e., computationally intractable (it means that there is no polynomial algorithm to find a solution in a reasonable time). An approach to deal with these problems is the use of metaheuristics algorithms. Researchers have used single objective optimization based models to deal with biological problems applying a wide number of metaheuristic techniques: simulated annealing, variable neighborhood search, particle swarm optimization, memetic algorithms, tabu search, evolutionary algorithms and other bioinspired techniques. However, the globally optimal solution is not guaranteed due to local minimum stagnation, lack of gradient and data noise influence. Biological optimization problems generally involve numerous objectives, which may consider different aspects of the solutions, which can be incommensurable and often, partially or wholly in conflict. Using multiobjective optimization formulation is possible to model these problems, decompose one objective function into more functions to reduce the probability of local minimum stagnation, add goals to solve problems with gradient and reduce noise using secondary bias function. However, finding a set of Pareto-optimal solutions and chosing only one solution using a decision maker increases the computational cost. We show a comparison between single and multi objective optimization based models in sequence alignment and protein structure prediction. The goal is to highlight the advantages on using a multiobjective approach despite the extra computational effort needed.

[1]Departamento de Ingeniería Informática, Universidad de Santiago de Chile, USACH

{manuel.villalobos, mario.inostroza@usach.cl}

A medical bioinformatics approach for metabolic inherited disorders associated with cancer: A case study on D2/L2-Hydroxyglutaric Aciduria related to Glioma with IDH1/2 deficiency

Vallejo-Ardila Dora L.[1,2], Ida Vanessa D. Schwartz[1,2], Fernanda SperbLudwig[1,2]

After Otto Warburg's work "On the Origin of Cancer Cells" was published in 1956, other studies of cancer metabolism have been reinforcing the concept that aberrant energetic metabolism occurs in cancer cells prior to the sequence of events that results in carcinogenesis. The availability of whole genome analyses data has facilitated the discovery of clinically relevant genetic alterations in cancer and inborn errors of metabolism, but still remains unclear what would be the potential role for those metabolic enzymes in cancer development. This fact provides us not only with the challenge to assess functional associations between an experimentally derived gene set of interest and a database of known gene sets, but also to integrate biological knowledgebase and analytic tools aiming at systematically exploiting biological meaning from large gene list. A case study on L2-hydroxyglutaric Aciduria related to Glioma with IDH1/2 deficiency is demonstrated by using network-based set enrichment analysis (EnrichNet) and modular enrichmen analysis (MEA). We applied a medical bioinformatics approach for comparing two types of methods of enrichment analysis to extract the biological insight into the pathogenesis of metabolic inherited disorders associated with cancer. Availability: EnrichNet: network-based gene set enrichment analysis is available at: http://www.enrichnet.org and The Database for Annotation, Visualization and Integrated Discovery (DAVID) is available at: http://david.abcc.ncifcrf.gov.

[1]Universidade Federal do Rio Grande do Sul (UFRGS)

[2]Hospital de Clínicas de Porto Alegre (HCPA)

A Knowledge-based Particle Swarm Optimization for the Protein Structure Prediction Problem

Mariel Barbachan e Siva[1]
Bruno Borguesan[2]
Márcio Dorn[2]

Proteins perform a wide range of functions in living organisms and consist of one or more chains of amino acids linked together by peptide bonds. This linear sequence is called the primary structure of the protein and tends to organize itself in regular conformations in space, these organization patterns are the secondary structure, the protein folding results in the tertiary structure which is closely related to the protein function. The three-dimensional protein structure prediction (PSP) is one of the most important problems of structural bioinformatics. Several computational strategies have been proposed for this problem, either using experimental information or not, the latter have a high computational cost associated. Metaheuristics are often used in attempts to solve the PSP problem due to their ability to find satisfactory solutions with lower computational effort than exact methods. Particle Swarm Optimization (PSO) is a metaheuristic of stochastic optimization inspired by the social behavior of animals. The algorithm works with a swarm of candidate solutions (particles) moving around the search space according to a fitness function. The movement of the particles is guided by their own best position in the search space and the best position of the swarm. This work is based on the fact that the native conformation of a protein corresponds to the one with the lowest potential energy. We developed a knowledge-based PSO for PSP problem. In our method, each particle is a candidate protein structure represented by a set of main and side chain torsion angles. Our knowledge-based swarm is generated by two different datasets. The first one have its structural information extracted from experimental protein templates by a clustering strategy combined with artificial neural networks and the second one from a database constructed by homology with protein fragments containing torsion angles information based on amino acid sequences and corresponding secondary structure. Preliminary results indicate that the approach using these datasets lead to better potential energy and RMSD results compared to swarms generated without this information.

Acknowledgements

This work was partially supported by grants from FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9).

[1] Instituto de Biociências, UFRGS
{mariel.barbachan@ufrgs.br}
[2] Instituto de Informática, UFRGS
{mdorn, bborguesan@inf.ufrgs.br}

Analysis of Protein Interaction in the Cancer Development

Luiz Henrique Rauber[1]
Cristhian AugustoBugs[2]
Éder Maiquel Simão[3]

The activation of Genome maintenance mechanisms (GMM) pathways such as: cell cycle (CC), DNA damage response (DDR) and apoptosis (APO) significantly contribute to tumor development. In previous studies, it was found in pre -cancerous activation process there is an anti-cancer barrier, which is responsible for prevention of tumor progression. The identification of the genes expressed during activation of the anti -cancer barrier, associated interactions in GMM pathways, becomes a complementarity to the study of the evolution of cancer. In this work, the objective was to investigate the activation of anti -cancer barrier in pre-cancer and cancer present in the adrenal gland tissue, colon, pancreas and thyroid follicles, using networks of interaction between proteins. To describe this barrier was proposed the modeling interaction networks between the proteins of MMG pathways using the Cytoscape software. The results obtained with the most prominent genes in expression and quantity of interactions were compared with the results of previous publications and reconfirmed the relevance of CDKN1A, CHEK1, ATR, TP53, MRE11A and XRCC4 genes. This analysis allowed the identification of other genes complementary to previous studies as SKP2, CCNO, FADD, RAD50, NBN, BIRC3, CDK2 and XRCC6 genes. These genes are associated and complement the studies on the activation of anti-cancer barrier. These considerations highlight that it is important to note the entire biological systemic and immersive context.

[1]Curso de Ciência da Computação, URI, CEP 97700-000
{Luiz Henrique Rauber, luiz.rauber@gmail.com}
[2]UNIPAMPA, CEP 97300-000
{Cristhian Augusto Bugs, cristhianbugs@gmail.com}
[3]Programa de Pós Graduação em Nanociências, UNIFRA, CEP 97010-491
{Éder Maiquel Simão , edersimao@gmail.com}

Analysis of Histones Deacetylases Expression and Post-transcriptional Regulation in Pancreatic Ductal Adenocarcinoma

Cleandra Gregório[1,2], Bárbara Alemar[1,2], Mariana R. Mendoza[3], Alessandro Osvaldt[4,5], Rudinei Correia[2], Raquel C. Rivero[5,6], Simone M. S. Machado[5], Patrícia Ashton-Prolla[1,2]

Pancreatic ductal adenocarcinoma (PDAC) is a highly lethal and aggressive disease. The disruption of histone acetylation through histones deacetylases (HDACs) and expression regulation by miRNAs can lead to tumor development. In this study we assessed HDAC1, HDAC2, HDAC3 and HDAC7 expression in PDAC and non-tumoral tissue samples using experimental and bioinformatics analysis, correlated their expression levels with clinicopathological features in patients and performed in silico investigation of regulation by miRNAs. Expression levels of HDAC1, 2, 3 and 7 were measured by qRT-PCR from 25 PDAC and 23 non-tumoral adjacent tissues and their association with clinicopathological parameters was analyzed. Differential expression (DE) and correlation analyses of HDACs and miRNAs in PDAC was performed using five GEO microarray datasets. Potential miRNA-HDACs relationships were collected from miRNA interaction databases. $P<0.05$ was considered statistically significant. We found reduced expression in PDAC compared with non-tumoral tissues for all HDACs analyzed, with $P<0.05$ for HDAC1, 2 and 3. However, fold-changes were very small and not biologically relevant. Strong positive correlation was observed between HDAC1 and HDAC3 ($P=0.003$), and moderate negative correlation between HDAC1 and HDAC7 ($P=0.017$) and HDAC3 and HDAC7 ($P=0.032$). None of HDACs expression was correlated with clinicopathological features. DE analysis suggested significant up-regulation of HDAC1, 2 and 7, and down-regulation of HDAC3, albeit all of them associated with small fold changes. 728 miRNAs (44 DE) were retrieved by bioinformatics analysis as HDACs regulators, and 125 predicted miRNA-HDAC pairs had negative expression correlation. Twenty miRNAs were DE and negatively correlated with their targets, thus representing promising regulatory mechanisms to be further investigated. Our results indicate that there may be a role of HDACs in the pathogenesis of this tumor, but differential expression between groups was subtle. Biologically relevant role for HDAC1, 2, 3 and 7 in pancreatic carcinogenesis is questionable.

Financial support: (Grant #559814/2009-7), CAPES and Fundo de Incentivo à Pesquisa-FIPE, Hospital de Clínicas de Porto Alegre (Grant 10-0162 and 11-0510)

[1]Programa de Pós-graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil. Corresponding author: {cleandra.gregorio@gmail.com}

[2] Laboratório de Medicina Genômica, Centro de Pesquisa Experimental - Hospital de Clínicas de Porto Alegre.

[3]Laboratório de Pesquisa Cardiovascular, Centro de Pesquisa Experimental - Hospital de Clínicas de Porto Alegre

[4] Grupo de Vias biliares e Pâncreas - Hospital de Clínicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil.

[5]Hospital de Clínicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil.

[6] Departamento de Patologia, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil.

Relational Database for Genetic Analysis

Timóteo Matthies Rico[1]
Andrea von Groll[1]
Karina dos Santos Machado[2]
Pedro Eduardo Almeida da Silva[1]

Biological data that are stored in public database are generally available in flat files. Among several types of this format, one of the most popular is the FASTA format. However, providing those data in flat files is often inadequate for complex queries. The softwares that analyse biological data in flat files format are limited when complex filters and crossinformation are needed. Therefore, this study aims the implementation of a relational database for effective organization and analysis of data related to genre, species, strains, genes, and nucleotide of any type of organism. Furthermore, we developed a software that insert biological data into the proposed database, based on multiple alignment FASTA files. The database was developed in the Database Management System PostgreSQL version 9.3 and the software was developed in Java platform. Sequences from Mycobacterium tuberculosis were downloaded in FASTA format and inserted into the database to test the database schema and the software's feature of data import. Results show the use of the developed database allows the extraction of different information as: gene relation per species/strains; distinct sequence per gene; which strains have the same sequence in determined gene. Besides, important queries as identifying conserved regions and the molecular markers were also developed. The database ensures data integrity, avoids redundancy, and ensures high performance, even in complex queries, using appropriately its features such as primary and foreign keys, unicity, and indexes in specific fields. As future work we propose the development of an interface to generate appropriate data mining input files using the stored data on the proposed database.

[1]Programa de Pós graduação em Ciências da Saúde, Universidade Federal do Rio Grande (FURG)
{timoteomr@gmail.com, avongrol@hotmail.com, pedrefurg@gmail.com}
[2]Centro de Ciências Computacionais, Universidade Federal do Rio Grande (FURG)
{karinaecomp@gmail.com}

Identification and Analysis of Transcription Factors in Rice Under Iron Overload

Artur Teixeira de Araujo Junior[1]; Daniel da Rosa Farias[2]; Railson Schreinert dos Santos[2]; Danyela de Cássia Oliveira da Silva[2]; Solange Ferreira da Silveira Silveira[2]; Camila Fernanda de Oliveira Junkes[3]; Luciano Carlos da Maia[2]; Antonio Costa de Oliveira[4]

1 Introduction

Rice (*Oryza sativa* L.) is the staple food for more than two thirds of the world´s population, and the second most widely grown cereal in the world. Irrigated rice is a crop with large importance in the state of Rio Grande do Sul (Brazil), accounting for two thirds of total production in the country. When rice is under conditions of iron excess, symptoms of toxicity are observed and may cause a decrease in productivity [1].

To respond and adapt to diverse abiotic stresses, as iron overload, many plant genes are induced or repressed, changing the levels of a variety of proteins. In this context transcription factors play an important role, since they do not only regulate growth and development, but also response to biotic and abiotic stresses [2]. These proteins have the capacity to recognize and bind to specific DNA sequences, regulating negatively or positively the transcription of a target gene [3]. Recently, several studies aiming to identify genes responsive to stress by iron toxicity have been described, however, in the literature there is a predominance of studies on iron deficiency. Due to little information about this topic, this work aimed to perform an analysis of these transcription factors. Here we managed to discover the transcription factor family which had its transcriptional activity altered more frequently in response to iron excess, predicted the protein-protein interaction of these genes and performed an analysis to detect conserved motifs in the promoter region of these genes.

2 Materials and Methods

According to the study of [4], differentially expressed genes under iron stress were selected. This genes where then characterized as transcription factors through Plant Transcription Factor Database (PlantTFDB) [5]. Their domains were identified in the Protein Families Database (Pfam) [6], based on the sequences predicted by Expert Protein Analysis System (ExPASy) [7]. Promoter phylogenetic analysis was performed using the sequences 1 kb upstream of each gene, using Molecular Evolutionary Genetics Analysis 6 - MEGA 6 [8] with ClustalW [9] and NeighborJoining [10] with 10,000 replicates of bootstrap. The identification of conserved motifs was performed using the same sequences in the Multiple Motif In Elicitation - MEME 4.4.0 [11], using a width ranging from 6 to 65 nucleotides using both strands, when analyzing promoter sequences, and 6 to 65 amino acids, when analyzing protein sequences, and a maximum of 10 motifs. Subsequently, these proteins were analyzed with the Predicted Rice Interactome Network – PRIN [12] and GENEVESTIGATOR [13] to verify protein-protein interactions and gene-expression analysis, respectively.

[1] Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900 (arturtaj@hotmail.com)
[2]Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900
[3]Universidade Federal do Rio Grande do Sul, UFRGS, Caixa Postal 110 CEP 90040-060
[4]Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900 (acostol@cgfufpel.org)

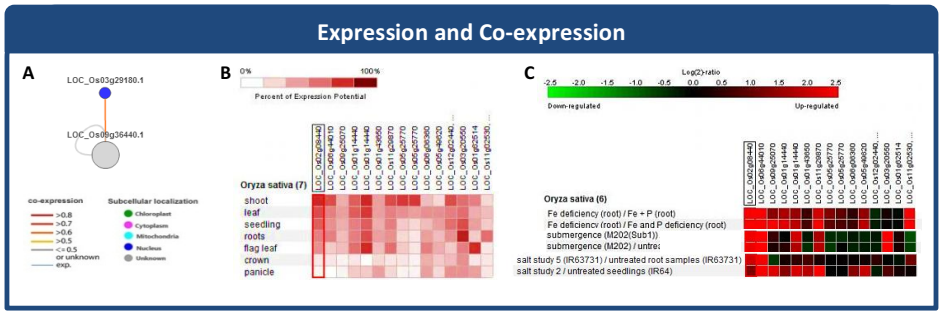
3 Results and Discussion

We identified the presence of 50 sequences characterized as transcription factors. Of these only one was down-regulated (Os02g0677300) while the others had an up-regulation. We analyzed the correlation of these proteins through PRIN, and found four recognized proteins, but only LOC_Os09g36440.1 presented correlation with another differentially expressed protein (LOC_Os03g29180.1). This co-expression value was of 0.6344. In this sense, the lack of information about protein-protein interaction on the sequences found in this work suggests a wide diversity, which is may not be well represented in the database.

We found 21 different domains characterized as transcription factors, wherein the most frequent was WRKY DNA-binding domain, with 15 occurrences, then the AP2 domain (6 occurrences) and Myb-like DNA-binding domain (5 occurrences). According to [14] the WRKY proteins are newly identified transcription factors involved in many plant processes, including plant responses to biotic and abiotic stresses. We analyzed the expression of these 15 WRKY genes in GENEVESTIGATOR, demonstrating that generally, when we observe the plant tissue we have many of these genes not expressed or have a low potential for expression in leaf, justifying this increase obtained in expression may have been promoted by stress conditions. For other stresses, we have iron deficiency stress with usually a high expression of these genes, the submergence stress has usually a down expression or no change in the expression of these genes and excess salt stress showed mostly a non-altered expression, due this we can see that the stress response mechanism which iron unleashed the expression of these genes following a different route response to other stresses.

The phylogenetic analysis of these 15 genes, formed three different groups when their amino acid sequence was analyzed, and other three groups when analyzing the nucleotide sequence of their promoter regions. The data obtained in classification of the amino acid sequences are confirmed by [14,15].

According to these results, we can conclude that changes in transcript accumulation of transcription factors are an important form of adaption to iron stress in *Oryza sativa* L. ssp. *japonica* cv. Nipponbare. There is still little information about iron stress in rice, and this work can enrich the databases aimed for breeding purposes. The identification of regulators of transcription can aid its use in genetic engineering and marker development.



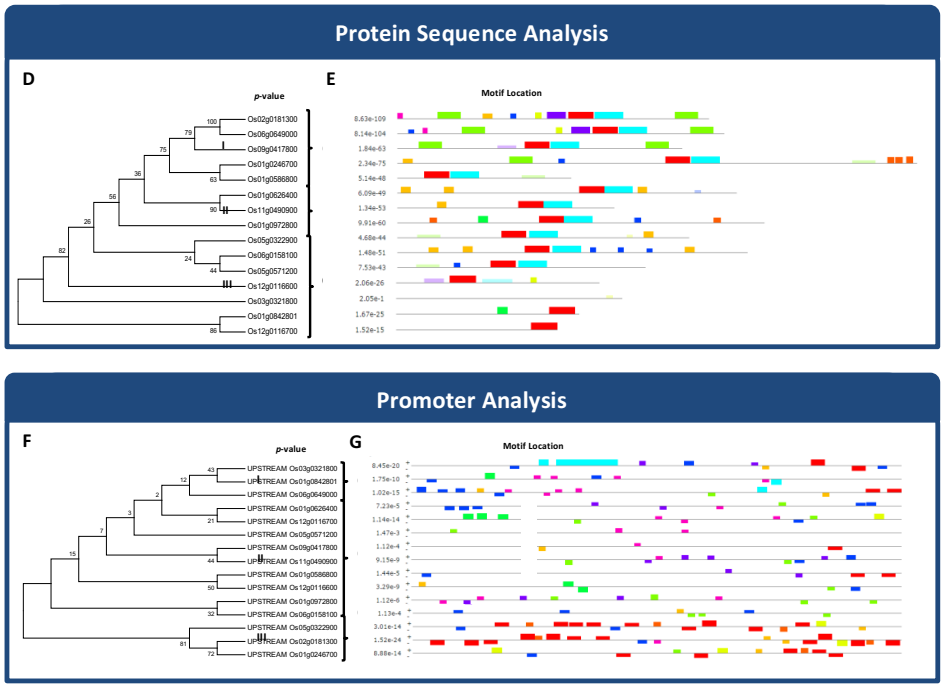


Figure 1 (A) Protein-protein interactions network. Expression analysis: (B) Localization and (C) Stress condition. (D) Phylogenetic tree and (E) Motifs found on WRKY proteins. (F) Phylogenetic tree and (G) Motifs found in the promoter region (1 kb upstream) of the analyzed WRKY genes.

References

[1] Pierre JL, Fontecave M. Iron and activated oxygen species in biology: the basic chemistry. Biometals. Sep;12(3):195-9. 1999.

[2] Zhu JK. Salt and drought stress signal transduction in plants. Annu Rev Plant Biol. 53:247–73. 2002.

[3] Krol R, Blom J, Winnebald J, Berhörster A, Barnett MJ, Goesmann A, Baumbach J and Becker A. RhizoRegNet—A database of rhizobial transcription factorsand regulatory networks. Journal of Biotechnology 155, 127– 134. 2011.

[4] Finatto T, Oliveira AC, Chaparro C, Maia LC, Farias DR, Woyann LG, Mistura CC, Soares-Bresolin AP, Llauro C, Panaud O and Picault N. Abiotic stress and genome dynamics: specific genes and transposable elements response to iron excess in rice. Rice. 8:13. 2015.

[5] Jin JP, Zhang H, Kong L, Gao G and Luo JC. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Research, 42(D1):D1182-D1187. 2014.

[6] BATEMAN A. et al. The Pfam protein families database. Nucleic Acids Research, Oxford, v. 30, p. 276-280, 2002.

[7] GASTEIGER E. et al. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research, Oxford, v. 31, p. 3784-3788, 2003.

[8] TAMURA, K. et al. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution, Oxford, v. 30, p. 2725–2729, 2013.

[9] THOMPSON J.D.; HIGGINS D.G.; GIBSON T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, Oxford, v. 22, p. 4673-4680, 1994.

[10] SAITOU N.; NEI M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, Oxford, v. 4, p. 406-425, 1987.

[11] BAILEY T.L. et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research, Oxford, p. W202-W208, 2009.

[12] GU H.B., ZHU P.C.; CHEN M. PRIN: a pedicted rice interactome network. BMC Bioinformatics, 2011.

[13] Zimmermann P, Hirsch-Hoffmann M, Hennig L, and W Gruissem GENEVESTIGATOR: Arabidopsis Microarray Database and Analysis Toolbox Plant Physiology 136 1 , 2621-2632. 2004.

[14] Zhang Y, Wang L. The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. BMC Evol Biol. Jan 3;5:1. 2005.

[15] Eulgem T, Rushton PJ, Robatzek S and Somssich IE. The WRKY superfamily of plant transcription factors. Trends Plant Sci. May;5(5):199-206. 2000.

A Distributed Knowledge-based Genetic Algorithm for Protein Structure Prediction

Jonas da S. Bohrer[1]
Bruno Borguesan[1]
Márcio Dorn[1]

The study of proteins and the prediction of their three-dimensional (3-D) structure is one of the most challenging problems in Structural Bioinformatics. Over the last years, several computational strategies have been proposed as a solution to this problem. As revealed by recent CASP experiments, the best results have been achieved by knowledge-based methods, but further research remains to be done. In this work, we propose a distributed knowledge-based Genetic Algorithm to predict the 3-D structure of proteins. A Genetic Algorithm (GA) is an adaptive heuristic search algorithm based on evolutionary ideas. GAs are modeled through the use of a population of individuals that undergo selection in the presence of variation-inducing operators such as mutation and recombination (crossover). A fitness function is used to evaluate individuals, and reproductive success varies with fitness. The success of the GA depends mainly on the balanced exploration of the solution space. When this balance is disproportionate, a premature convergence problem can occur, and the GA will lose efficacy. In protein structure prediction, the roughness of the protein energy surface poses a significant challenge to optimization techniques such as GAs. One approach to deal with this problem considers the use of a Distributed Genetic Algorithm. The basic idea of a distributed GA is to keep, in parallel, independent populations. We proposed a Distributed Genetic Algorithm based on the conformational preference of amino acid residues in experimentally-determined proteins. Each population incorporates an Angle Probability List (APL) derived from experimental data to generate the initial populations and new individuals to increase the diversity of the model. The proposed method has been tested with eight protein sequences. As corroborated by experiments, the developed method can produce accurate predictions, where the 3-D protein structures are comparable to their corresponding experimental ones. When compared with other first principle prediction methods that use database information, our approach presents advantages in terms of demanded time to produced native-like 3-D structures of proteins.

[1]Instituto de Informática, UFRGS, Porto Alegre, RS, Brazil.

{jonas.silveirabohrer,bborguesan,mdorn@inf.ufrgs.br}

OncoProSim: A tool for in silico tumor evolution analysis

Darlan Conterno Minussi[1,2]
Bernardo Henz[4]
Eduardo Cremonese Filippi-Chiela[3]
Manuel Menezes de Oliveira[4]
Guido Lenz[1,2]

The current knowledge regarding tumor evolution, until now, has been obtained from what we can gather through biopsy samples. In spite of the importance of these samples to diagnosis and treatment, they represent only a glimpse of the whole tumor evolutionary path, whereas the majority of the tumor development remains hidden. In this work, we attempt to simulate the clonal evolution of different types of tumors, focusing on the molecular mechanisms that lead to tumor progression. In order to do that, we generate cells with two arrays that represent the diploid characteristic of the human genome and, in case of cell division, mutations can be inserted in the genome that may alter the default probabilities of proliferation and death. With the help of a pseudorandom number generator, we can use our model to simulate different patterns of tumor evolution and investigate the effects of different mutational spectrum in the incidence of distinct tumors. Moreover, we can use our model to reproduce essential characteristics of the tumor genome such as: the synergy between oncogenes and tumor suppressor genes, changes in mutation frequency with distinct fitness values and even simulate tumor treatment. Therefore, with the increasing knowledge in tumor biology, we believe that our model can offer a unique perspective of tumor evolution, allied to the speed and reproducibility that only in silico models are capable of offering.

[1]Programa de Pós-Graduação em Biologia Celular e Molecular, UFRGS (darlanminussi@gmail.com)
[2]Instituto de Biociências, Departamento de Biofísica, UFRGS
[3]Faculdade de Medicina, UFRGS
[4]Instituto de Informática, UFRGS

Side-Chain conformational analysis of the multi-dependent rotamer preferences of proteins

Bruno Borguesan[1]
Mariel Barbachan e Silva[2]
Márcio Dorn[1]

Rotamer libraries are commonly used to correctly assign the dihedral side-chain angles for amino acid residues in protein structures. Rotamer libraries are widely used to assist the problems of Structural Bioinformatics like the protein structure prediction, protein design, structure refinement, homology modeling, and X-ray and NMR structure validation. These libraries are mostly classified as backbone-dependent, backbone-independent and secondarystructure-dependent. The first group are libraries that consists of rotamer frequencies, mean dihedral angles, and variances as a function of the backbone dihedral angles. The second group makes no reference to backbone conformation, but use side-chain information from all experimentally determined protein structures available. The third group present rotamer frequencies and dihedral angles for each secondary structures, such as α-helix, β-sheet and coil. However, even when these romater libraries are applied an enormous possibility of sidechain angles can be allowed. To reduce the complexity of the side-chain search within the Tertiary Protein Structure Prediction problem, we propose a novel approach that combine backbone-dependent, amino-acid-dependent and secondary-structure-dependent. To develop this method, we performed a study with 6,650 protein structures to analyze the conformational preferences for the side-chain angles of each amino acid residue, computed with PROMOTIF, in the secondary structure assigned by STRIDE. From that point forward, we combined the side-chain angles of each amino acid residue in the secondary structure with the frequencies of backbone dihedral angles. Based on this analysis, we developed a rotamer library that combine all these conformational preferences to assign the dihedral side-chain angles for amino acid residues in protein structures. This multi-dependent rotamer library was already used to assist implementations with PyRosetta in tertiary structure prediction.

This work was partially supported by grants from FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9).

[1]Institute of Informatics, UFRGS, Av. Bento Gonçalves 9500, 91501-970 Porto Alegre, RS, Brazil{bborguesan,mdorn@inf.ufrgs.br}

[2]Institute of Biosciences, UFRGS, Av. Bento Gonçalves 9500, 91501-970 Porto Alegre, RS, Brazil {mariel.barbachan@ufrgs.br}

A Multi-Agent Approach for the 3-D Protein Structure Prediction Problem

Leonardo de L. Corrêa[1]
Márcio Dorn[1]

The prediction of the three-dimensional structure of proteins is an important research area in Structural Bioinformatics. Proteins are present in all living systems, performing different functions. The function performed by a protein is strictly related to its adopted conformation. This is the main motivation for researchers in the field, besides the large gap between the number of known protein sequences and known 3-D protein structures. The protein structure prediction (PSP) problem is classified in computational theory as a NP-complete problem due the high dimensionality and complexity that the search space can assume, even for a small protein. Therefore, there is not any method capable to achieve the optimal solution. Currently, to predict the 3-D structure of a protein only from the amino acids sequence (ab initio methods), a wide range of optimization algorithms are being applied to find approximated solutions, as well as previous knowledge of known protein structures stored in a protein data bank in order to improve these methods and reduce the search space. In this work, we propose a multi-agent system to deal with the PSP problem. Multiagent systems are being used to face complex problems, devising tasks among the agents of the system and exploring a more distributed approach. An agent is a computer process that runs in an independent way, where under some given circumstances, interacts and cooperates with the other agents to solve a major problem. In our approach, we employed a population of thirteen agents that were organized in a structured ternary tree, whereas each agent performs specific roles and interacts in an attempt to obtain good solutions to the problem. We also incorporate, as behaviors of the agents, concepts of evolutionary-based algorithms, such as crossover and swap operators, and a Simulated Annealing implementation as a local search method. The algorithm also uses the information stored in the Protein Data Bank through an Angle Probability List (APL) to reduce the search space. The APL was partitioned into sub-groups, according to the combination of amino acids residues and their respective secondary structures, resulting in different patterns. The population of agents was divided into subpopulations, to increase the diversity in the space of solutions, each agent could see only one pattern of the APL. We used the Rosetta energy function as scoring function. The system was tested against three protein sequences, where for each one the algorithm was run four times for eight hours. Preliminary results show that the proposed approach has a promising ability to predict good three-dimensional conformations in terms of potential energy and RMSD measurement when compared with the experimental protein structures.

This work was partially supported by grants from FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9)

[1]Institute of Informatics, Federal University of Rio Grande do Sul, Post Code 24105
{llcorrea, mdorn@inf.ufrgs.br}

Effects of Zn^{+2}, N-Glycosylation and Dimerization on Auxin Binding Protein 1 (ABP1) Conformation and its Interaction with Auxin

Cibele Tesser da Costa[1], Conrado Pedebos[1], Hugo Verli[1], Arthur Germano Fett-Neto[1]

Auxin is a critical phytohormone for plant growth and development, influencing aspects of cell division, elongation and differentiation. One natural form of auxin is the indole-3-acetic acid (IAA), whereas 1-naphtalene acetic acid (NAA) is a synthetic form. Auxin Binding Protein 1 (ABP1) is an auxin receptor that plays key roles in fast nontranscriptional responses. Its structure was previously established from maize at 1.9 Å resolution and each subunit of ABP1 is glycosylated through a high mannose-type glycan structure. In dicot species it has an additional glycosylation site. We aimed at expanding the knowledge of ABP1 structural biology employing molecular dynamics simulations of the complete models of the oligomeric glycosylated proteins for maize and Arabidopsis thaliana with or without auxins. In maize, both coordination to Zn^{+2} and glycosylation were able to promote increased conformational stability whereas in Arabidopsis we observed only a minor effect of both factors. Most of such stabilization effect was located on N- and Cterminal regions. C-terminal regions of ABP1 from both species contain an α-helix and in the performed simulations this helix unfolded, assuming a more extended structure in all replica simulations of maize ABP1. In Arabidopsis the helix seems to be more stable, beingpreserved in most of the monomeric simulations and unfolding when the protein was in the dimeric form. When Arabidopsis ABP1 was bound to IAA or NAA, the glycosylation structures arranged around the proteins, covering the putative site of entrance or egress of the ligand. Along the simulation, when NAA was bound to the proteins, the fold was more similar to the crystal structure and the complex showed higher stability compared to IAA binding. This work contributed to a better understanding of the effects of Zn^{+2}, glycosylation and dimerization in the structure of the protein ABP1 and its interaction with auxins.

Support: Capes, CNPq and Fapergs

[1]Center for Biotechnology, Federal University of Rio Grande do Sul (UFRGS), CP 15005, Porto Alegre, RS, 91501-970, Brazil

Memetic algorithm for Docking's problem with a rigid protein and a flexible ligand

Felipe Gonzalez Foncea[1]
Mario Inostroza-Ponta[1]
Márcio Dorn[2]

Docking is a problem of bioinformatics, which has achieved acceptance over the past 20 years, due to advances in the field of molecular biology. The search for new structures has defined the Docking as a viable alternative for finding an optimal bonding between two molecules, mainly a protein and a ligand. Protein-ligand structure can generate alterations in signal transduction process, which is a key factor in the chemical processes of a biological body. This knowledge will validate the creation or discovery of new types of drugs. As a research proposal, It has been worked to generate a new alternative for Docking problema, through a memetic algorithm and the Rosetta software suite as a scoring function, which evaluates the level of protein-ligand affinity through free energy binding. The model is restricted to use proteins without any change on the structure during all the evaluation process. However ligand will be able to change their structure to generate more alternatives of solution. The flexibility of the protein structure will be produced by the algorithm, that changes the dihedral angles of the ligand. The results will be evaluated by the scoring function. The choice of a rigid protein and a flexible ligand is focused on getting a good result without high computing power. In addition it would be possible to use this model as a local search for more complex models both protein and ligand flexible. The objective of this algorithm is solve most of the problems associated to Docking. The results will be analyzed by data quality through RMSD (Root-Mean-Square Deviation) and it will be compared with other Docking programs like AutoDock and Dockthor.

[1] Departamento de Ingeniería Informática, Universidad de Santiago de Chile
{felipe.gonzalez.f@usach.cl, mario.inostroza@usach.cl}
[2] Institute of Informatics, Federal University of Rio Grande do Sul
{mdorn@inf.ufrgs.br}

Análises de elementos transponíveis em *Angiostrongylus cantonensis* (*Nemathelminthes: Angiostrongylidae*)

Alice Giovana Buzetto[1]
Leandro de Mattos Pereira[1]
Gabriel da Luz Wallau[1]
Amaranta Ramos Rangel[1]
Alessandra Loureiro Morassutti[1]
Carlos Graeff-Teixeira[1]

Tansposons são elementos móveis do genoma que se fazem presentes em todos os organismos, correspondendo até 50% do genoma em algumas espécies. Pela sua natureza móvel, possuem grande importância genômica, sendo fontes de mutações espontâneas, rearranjos cromossômicos, produzem efeitos na evolução das espécies e podem ser utilizados como marcadores moleculares. O estudo dos transposons em uma espécie de nematódeo com importância em saúde pública, *Angiostrongylus cantonensis*, tem como objetivo identificar a presença destes elementos móveis a fim de compreender seu papel na evolução da espécie e a nível molecular, bem como sua possível relação com os mecanismos de patogenicidade. Foram utilizados dois métodos de análises para a identificação dos transposons: análise por homologia (TBLASTN) e *ab initio* e homologia (RepeatExplorer). A análise baseada em homologia foi aplicada utilizando o algoritmo TBLASTN, a qual consistiu em comparar as sequências de transposons de *Caenorhabditis elegans* com sequências genômicas de *A. cantonensis*. Enquanto que a análise híbrida (RepeatExplorer), executa uma análise global no genoma de interesse. Os dados preliminares apresentam 1.149 sequências de elementos repetitivos, obtidas na análise por TBLASTN, curadas no banco de dados RepBase, utilizando o algoritmo CENSOR. Destas sequências, observam-se retrotransposons e transposons, classificados em classe e família, tendo destaque a família de transposon Mariner/Tc1, por compreender maior número de repetições. As análises com o programa RepeatExplorer revelam 124 clusters, compreendendo sequências repetitivas de retrotransposon e transposon. Os resultados obtidos até o momento, associados a dados já publicados com outros organismos, sugerem a participação destes elementos na evolução das espécies, uma vez que identificam-se alguns transposons de mesma classe e família em diversos organismos, previamente descritos e depositados em banco de dados.

[1] Laboratório de Biologia Parasitária - Pontifícia Universidade Católica do Rio Grande do Sul

A Python-Based API of the 3D-Tree Algorithm

Aline Kronbauer[1] ;Bruna L. Balbinot[1], Leonardo A. Schmidt[1], Leonardo N. Machado[1], Raphael G. Nascimento e Silva[1], Karina S. Machado[2], Ana T. Winck[1]

Rational Drug Design is an important field of bioinformatics, treating mainly about interaction between macromolecules, called as receptors, and small molecules, called as ligands. By means of molecular docking and the estimated Free Energy of Binding (FEB) we can identify the best bind between the molecules. Many studies consider the receptor as a rigid structure, ignoring its flexibility. Aiming at reaching better results we make use of several molecular docking experiments, being that each of them run on a given receptor conformation generated by molecular dynamic (MD) simulations. However, we know that such strategy generates a lot of data, turning it a timing-consuming process. Different methods have been developed to treat this issue, and some of them try to select the most promising conformations from the whole MD. In this work we focus on the 3D-Tree algorithm, which identify the best pose of relevant receptorâs atoms that leads to a good estimated FEB value. We are implementing a Python-based version of the 3D-Tree algorithm, aiming to turn available an API to provide the whole process of this algorithm. In summary, the 3D-Tree algorithm treats spatial coordinates in a x, y, z format, and induce a decision-tree that predicts FEB. The algorithm uses the spatial coordinates to binary split a node, where the edges evaluate whether the atom is part of a given block. Hence, we develop a full preprocessing step, where a set of PDB files can be uploaded to generate a dataset fitting the 3D-Tree file format. Having these dataset, the API is able to submit them to the next two main steps of the algorithm, which regards the block generation and the induction decision-tree step. Currently, we are testing our implementation on the AcrB protein, which contains 7.639 atoms and 1.001 PDB files obtained by MD simulations.

Acknowledgments: This work was supported by grants from MCT/CNPq 14/2013 to ATW and KSM. AK is supported by a FAPERGS scholarship. The 3D-Tree algorithm is a PhD theses developed at PUCRS from 2008 to 2012.

[1]Grupo de Pesquisa em Sistemas Inteligentes - UFSM
alyne.k@gmail.com,{bbalbinot,lschmidt,lmachado,rnascimento,ana}@inf.ufsm.br
[2] Grupo de Pesquisa em Biologia Computacional - FURG
karina.machado@furg.br

Data Mining Applied to Molecular Docking Experiments

Luisa Rodrigues Cornetet[1]
Michael González-Durruthy[2]
José M. Monserrat[2]
Adriano V. Werhli[3]
Karina dos Santos Machado[3]

Introduction

Since Bioinformatics experiments as molecular docking and virtual screening usually produce a lot of data, it is really important to apply data mining techniques to discover useful information about this generated data. In this paper, we propose to apply data mining classification algorithms into the results of molecular docking experiments with carbon nanotubes and the protein ANT [ant].

Molecular docking is a computer technique that helps to understand the interaction between biological macromolecules, called receptors, and some small molecule that are inhibitor candidates, called ligand. The analysis of molecular docking allows to understand if a ligand acts as an inhibitor or antagonist during a physiological process [bioinformatica]. For the molecular docking, we perform the experiments using Autodock Vina [vina]. Autodock Vina is a open-source software developed by Trott et al [vina]. As the results of the molecular docking simulation we obtain the conformations of the ligand and the Free Energy of Binding (FEB) of the receptor-ligand complex. Also, we used a framework [luisa] to make the execution easier and calculate the minimum distances between atoms of the aminoacid and the ligand on the results.

Metodology

In this section we will present the receptor and the ligands used in the docking experiments, as well as the pre-processing of the data for the data mining process.

As receptor in the molecular docking experiments we consider the protein ANT [ant], or Adenine Nucleotide Translocator. This protein is located on the inner wall of mitochondria and is an essential part on the process of energy production for a eukaryotic cell. ANT acts as a carrier, and performs the exchange between ADP and ATP. This exchange works with importing ADP and exporting ATP across the inner membrane of mitochondria. The structure of this protein was obtained from Protein Data Bank (PDB) [pdb] with the PDB ID 1OKC.

1 Centro de Ciências Computacionais, FURG.
{cornetet.luisa@gmail.com}
2 Instituto de Ciências Biológicas, FURG.
3 Centro de Ciências Computacionais, FURG.

As ligand we have considered several types of carbon nanotubes. The carbon nanotubes are a material made of a graphene sheet in a tube form, and the diameter in nanometric scale [SWCNT]. In this work we have varied parameters such as the Hamada Index and the oxidation of the carbon nanotube. All the experiments were performed with single wall carbon nanotubes (SWCNT). The Hamada Index [hamada] is expressed by two parameters,n and m, that classify the carbon nanotubes for its diameter and helicity. As de n and m parameter vary, we have three classifications for this nanotubes: armchair (when m = n), chiral (when m 6 = n and m = 0) and zigzag (when m = 0). We have used 7 armchair nanotubes, 31 chiral nanotubes and 7 zigzag nanotubes. About the oxidation, we used pristine nanotubes, and also nanotubes with the oxidation groups carboxyl (COOH) and hydroxyl (OH), in all the previous described nanotubes.
As ligand we have considered several types of carbon nanotubes. The carbon nanotubes are a material made of a graphene sheet in a tube form, and the diameter in nanometric scale [SWCNT]. In this work we have varied parameters such as the Hamada Index and the oxidation of the carbon nanotube. All the experiments were performed with single wall carbon nanotubes (SWCNT). The Hamada Index [hamada] is expressed by two parameters, n and m, that classify the carbon nanotubes for its diameter and helicity. As de n and m parameter vary, we have three classifications for this nanotubes: armchair (when m = n), chiral (when m 6 = n and m = 0) and zigzag (when m = 0). We have used 7 armchair nanotubes, 31 chiral nanotubes and 7 zigzag nanotubes. About the oxidation, we used pristine nanotubes, and also nanotubes with the oxidation groups carboxyl (COOH) and hydroxyl (OH), in all the previous described nanotubes.

Results and discussion

In this section we describe and analyze the data mining techniques applied to the results of the molecular docking experiments. For the data mining algorithms we used Weka [wekaUpdate], that provides an environment for some techniques as automatic classification, clustering and regression [weka].
We applied J48 [j48] and REPTree [reptree] classification algorithms. In both cases we validate the obtained models with a 10-fold cross validation. Figure 1 shows the decision tree generated by J48 algorithm. This model had accuracy of 73.88%. The root node of the tree had divided the instances according to the Oxidation. According to Figure 1 we can observe that distances between the aminoacid Arg 234 and the nanotubes greater than 13.71Å produces better results, we believe it occurs because this aminoacid is a little lower in the protein, and the nanotubes were in a position above.

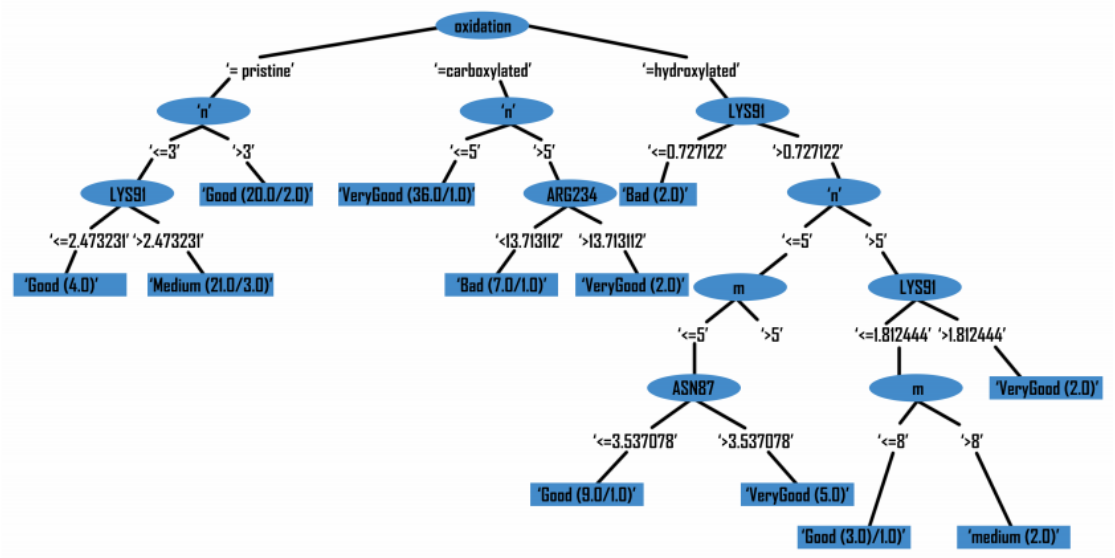


Figure 1. J48 results

Figure 2 shows the decision tree created by REPTree algorithm. It correctly classified 72.38% of the instances. As well as J48 decision tree, the root node of the tree is Oxidation. First of all, it says that every carboxylated nanotube has a Very Good FEB result, since the best experiment results are related to nanotubes that has the carboxyl group. Besides Good or Bad FEB results for pristine nanotubes is related to its size: the biggest ones ($n \geq 3.5$) has Good FEB values, while the small ones ($n < 3.5$) were classified as Medium. Finally the hydroxylated nanotubes presented Medium results for the ones with $n \geq 6.5$, and Good or Very Good for the others. It is also observed that very small distances between Lys 91 (less than 2.43) and the nanotubes obtained docking results with Bad FEB values and bigger values of distance between this aminoacid produces a Good result. We can explain this because very small distances between Lys 91 and the nanotubes might produce some collision of atoms.

Conclusion

Based on the presented results, we conclude that applying data mining techniques is very effective to discover important information about the ANT-Nanotubes molecular docking experiments. The results gave us a different point of view about the interactions between this target protein and different carbon nanotubes, allowing us to do further research about the results of the experiments.

Analysis of the metabolic potential of Angiostrongylus cantonensis

Amaranta Ramos Rangel[1], Leandro de Mattos Pereira[1], Alice Giovana Buzetto[1], Alessandra Loureiro Morassuti[1], Carlos Graeff-Teixeira[1]

Angiostrongylus cantonensis, the rat lungworm, is an etiological agent of eosinophilic meningitis in humans, has as hosts: Veronicellidae (intermediate), Rattus norvegicus (definitive) and Homo sapiens (accidental). This species has been founded in Southeast Asia, Pacific Basin and South of the Brazil (Porto Alegre). Currently, little is known about the metabolic pathways and biologic processes present in the genome and transcriptome of Angiostrongylus. Genome-wide analyses using computational biology are of fundamental importance for elucidating molecular mechanisms, metabolic pathways and molecular marker to diagnostics. From a holistic approach, the aim of this study was to evaluate the metabolic potential of A. cantonensis through of annotation of all metabolic pathways present in the genome and transcriptome of this parasite. We obtained the predicted proteome in the database WormDB and the transcriptome was sequenced by Macrogen with the Illumina platform. The Trinity software was used to de novo assembled of transcriptome and Trinotate, Blast2GO for functional annotation. Through of the AnEnPi we mapping all metabolic pathways present in the genome and transcriptome of A. cantonensis, including possible alternative routes. Our preliminary results shows that the predicted proteome (11,998 sequences) consists of 634 enzymatic activities, with 136 Oxidoreductases (Ecs1), 226 Transferases (ECS2), 193 hydrolases (Ecs3), 34 Lyases (Ecs4), 28 Isomerases (Ecs5) ligases (Ecs5), totaling 33% of the proteome. Among the metabolic pathways, we found metabolic pathways important for the survival of the parasite, such as: drug metabolism, carbohydrates metabolism and synthesis of nitrogenous bases.

Supported by Capes, Brazil. Acknowledge: Laboratório de Biologia Parasitária, PUCRS; Programa de Pósgraduação em Biologia Celular e Molecular, PUCRS.

[1]Laboratório de Biologia Parasitária- PUCRS
Pontifícia Universidade Católica do Rio Grande do Sul.

Development of a Genetic Algorithm for Protein-Ligand Redocking

Eduardo Spieler de Oliveira[1]
Márcio Dorn[1]

Molecular docking problems refer to the prediction of the conformation between a small molecule (ligand) and a receptor molecule with minimum binding energy. The quality of this binding depends on the score function and on the computation methods applied. The redocking method is often performed to verify if the docking parameters are able to recover a known complex structure and interactions. The development of docking methods include the use of metaheuristics to improve prediction capability. In this work, a Genetic Algorithm (GA) was developed in order to search efficiently the conformational space of the Protein-Ligand and find the minimum binding energy redocked structure. As a first study both structures were considered rigid and Rosetta score function was used to calculate the energy of the docking structures. Random perturbations of positions, such as translations and rotations, called individuals, were made in the ligand to search the ideal biding location. A population of one hundred individuals was created. The algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm randomly selects individuals from the current population and uses them as parents to produce the offspring for the next generation. The structure 1ENY from the Protein Data Bank was used as a reference. The algorithm ran fives times, each one for eight hours, over three hundred generations were tested toward an optimal solution. The final structure was compared with the original crystallographic structure in therms of root-mean-square deviation. Preliminary results show that the proposed algorithm can find good solutions in terms of RMSD when compared with the experimental crystallographic 3D structure.

Acknowledgements: This work was partially supported by grants from FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9).

[1]Instituto de Informática, UFRGS, Porto Alegre, RS, Brazil. {esolivera,mdorn@inf.ufrgs.br}