em questão
E-ISSN 1808-5245

Recuperação da informação e a consulta à base de dados no processo de busca do Mecanismo Online para Referências

Proxério Manoel Felisberto
Mestrando; Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil; proxerio.felisberto@posgrad.ufsc.br

Roderval Marcelino
Doutor; Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil; roderval.marcelino@ufsc.br

Alexandre Leopoldo Gonçalves
Doutor; Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil; a.l.goncalves@ufsc.br

João Bosco da Mota Alves
Doutor; Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil; jbosco@inf.ufsc.br

Resumo: Os dados das organizações crescem, exponencialmente, a cada ano e têm trazido aos administradores e gerentes incremento para tomada de decisão à qual são diariamente submetidos. Para a gestão destes dados, bem como, para descoberta de informações neles contidas, surgiram os Sistemas de Recuperação da Informação - largamente empregados no ambiente organizacional, atualmente. A Recuperação da Informação foi desenvolvida com a finalidade precípua de fornecer, rapidamente, aos usuários a informação que procuram. A avaliação de um Sistema de Recuperação da Informação é focada em seu motor de busca, medindo o quão rápido ele pode responder a uma consulta, ou o nível de relevância da informação recuperada. Este trabalho tem como objetivo verificar o impacto da utilização de motores de busca, baseados no *Apache Solr*®, no processo de recuperação da informação contida na base de dados do Mecanismo Online para Referências. Assim, buscaram-se na bibliografia, fundamentos para conceituar a Recuperação da Informação e tratar sobre as peculiaridades que se coadunam com o escopo desta pesquisa. Abordam-se as principais características do servidor de recuperação da informação *Apache Solr*® e do protótipo desenvolvido para os propósitos deste trabalho. Cabe esclarecer que o *Apache Solr*® foi configurado para ordenar os resultados pelo nível de relevância, sendo o Modelo de Espaço Vetorial utilizado no cálculo do grau de similaridade. Na sequência, os dados colhidos são tabulados, apresentados e analisados. Conclui-se que a utilização de motores de busca, baseados no *Apache Solr*®, impacta, positivamente, no processo de recuperação da informação contida na base de dados do Mecanismo Online para Referências.

Palavras-chave: Recuperação da informação. Indexação. Modelo de espaço vetorial. Mecanismo online para referências.

em questão
E-ISSN 1808-5245

1 Introdução

O volume de dados das organizações, em geral, tem crescido exponencialmente, nos últimos anos e tem trazido aos administradores e gerentes incremento na complexidade da tomada de decisão à qual são, diariamente, submetidos. Das decisões tomadas e implementadas na sua rotina diária depende o futuro da organização (WHITE, 2013).

Para auxiliar no gerenciamento destes dados, bem como na descoberta de informações neles contidas, passou-se a utilizar os Sistemas de Recuperação da Informação (SRI), que são largamente empregados no ambiente organizacional contemporâneo.

Acompanhando esta tendência, a base de dados do Mecanismo Online para Referências (MORE) tem apresentado significativo aumento no seu volume de dados, sugerindo que sejam buscadas alternativas para melhorar o tempo de resposta às pesquisas realizadas por seus usuários, uma vez que o processo de pesquisa à base de dados do MORE tem mostrado relativo aumento no tempo de resposta. Vale esclarecer que há três anos a base de dados do MORE armazenava um milhão de referências geradas, e o tempo de resposta a uma consulta oscilava em torno de três segundos. Ocorre que, neste período, o volume de dados armazenados aumentou seis vezes (atualmente, armazena quase sete milhões de referências), e o tempo de resposta passou a oscilar em torno de dezessete segundos, como pode ser observado nas Tabelas 1 e 2.

O MORE – desenvolvido em 2005 – é fruto da percepção da Sra. Maria Bernadete Martins Alves em relação às dificuldades encontradas pelos usuários da norma técnica referente à elaboração de referências bibliográficas no ambiente da Biblioteca Universitária da Universidade Federal de Santa Catarina (UFSC). Trata-se de uma ferramenta *web* gratuita destinada ao auxílio de usuários de bibliotecas (pesquisadores, professores e alunos) na tarefa de gerar e gerenciar suas listas de referências bibliográficas, de acordo com a norma NBR 6023:2002 da Associação Brasileira de Normas Técnicas (ABNT) e mantê-las armazenadas em uma base de dados própria (ALVES; MENDES; ALVES, 2006).

Atualmente, o MORE foi institucionalizado pela UFSC e encontra-se hospedado nos servidores da Superintendência de Governança Eletrônica e Tecnologia da Informação e Comunicação (SeTIC) daquela Universidade. Conta com mais de 160.000 usuários cadastrados e quase sete milhões de referências armazenadas em sua base de dados.

Os sistemas baseados na Web 2.0[1], além da interatividade, requerem constante atualização. Para acompanhar a evolução tecnológica empregada na recuperação da informação e estar preparado para a implementação de funcionalidades que se fizerem necessárias é que se buscam alternativas para cumprir os objetivos da recuperação da informação no contexto do MORE.

Xavier (2009) faz uma distinção entre sistemas de recuperação de dados e recuperação da informação, cuja diferenciação é observada pela utilização da técnica empregada para comparar a expressão de busca contra o repositório informacional, denominadas como equiparação exata e equiparação parcial, respectivamente. Percebe-se que o MORE faz uso da recuperação de dados. Esta afirmativa é baseada no fato do usuário informar sua expressão de busca, e o sistema, através de uma consulta SQL contra a base de dados, retornar somente os registros que expressam, completamente, a sequência de caracteres informada (equiparação exata).

Assim, identifica-se a oportunidade de implementação no sistema de busca do MORE da equiparação parcial, cujo método, segundo Xavier (2009, p. 25), "[...] possibilita tanto a obtenção de resultados parcialmente válidos ou pertinentes como a ordenação destes resultados de acordo com o grau de relevância para a expressão de busca utilizada.".

Poucos são os projetos desenvolvidos para sistemas de busca que utilizam a equiparação parcial e que sejam gratuitos. Deixa-se de considerar os sistemas proprietários pelo motivo de que o desenvolvimento do MORE é fundamentado na inclusão social e, por consequência, em tecnologias gratuitas. Verifica-se com frequência a existência de comparações entre o *Apache Solr*® e o *Apache ElasticSearch*® (ambos derivados do *Apache Lucene*®). O primeiro possui uma interface mais amigável para hospedagem em servidor único (que é o caso do MORE), mesmo havendo a possibilidade de sua clusterização (distribuição), enquanto o

segundo é mais voltado para bases de dados distribuídas (envolvendo múltiplos nodos). Do exposto, optou-se por utilizar o servidor de recuperação da informação *Apache Solr*® para dar suporte às consultas realizadas à base de dados do MORE.

Partindo do pressuposto que a utilização de recursos de Tecnologias de Informação e Comunicação (TIC) que satisfaçam o conceito de *Enterprise Search*[2] pode impactar positivamente no processo de recuperação da informação contida na base de dados do MORE busca-se, durante o desenvolvimento deste trabalho, verificar o impacto da utilização de motores de busca, baseados no *Apache Solr*®, no processo de recuperação da informação contida na base de dados em estudo.

Este artigo está estruturado em seis seções distintas, além das referências bibliográficas, por meio das quais pretende-se alcançar o objetivo proposto no parágrafo anterior.

2 Recuperação da informação

Esta seção dedica-se ao entendimento do conceito de recuperação da informação, visando facilitar a análise das necessidades requeridas pelo subsistema de busca do MORE e subsidiar a elaboração de uma estratégia de busca.

Segundo Lopes (2002, p. 61), no contexto da recuperação da informação, "[...] a estratégia de busca pode ser definida como uma técnica ou conjunto de regras para tornar possível o encontro entre uma pergunta formulada e a informação armazenada em uma base de dados.". Ou seja, selecionar e retornar ao interessado um conjunto de documentos (informações) que compõem a resposta à pergunta submetida à apreciação.

A história da humanidade e a evolução da ciência se confundem com a necessidade de recuperação e disponibilização de informações, que armazenadas em algum meio físico, possam ser resgatadas de forma precisa, rápida e com certo grau de relevância (FACHIN, 2009).

Classificar a informação da humanidade foi e tem sido requisito para organizá-la visando sua recuperação em tempo futuro. Para acompanhar o incremento da produção da informação foram desenvolvidos métodos, técnicas e

sistemas que almejam suprir os anseios referentes à sua recuperação que é fortemente afetada pela sobrecarga informacional (CARVALHO; LUCAS; GONÇALVES, 2010; OLIVEIRA; ARAUJO, 2012; TEIXEIRA; SCHIEL, 1997).

Para Ribeiro (2013, p. 531),

> [...] a classificação é assumida como uma operação intelectual e técnica, que se traduz numa categorização/sistematização para fins organizativos e numa representação formal tendo em vista a recuperação da informação.

Nesse sentido, percebe-se que as bibliotecas são pioneiras na valorização do acesso à informação, cuja assertiva advém da preocupação desta com a criação de dispositivos classificatórios, em benefício da representação e da recuperação da informação (TEIXEIRA; SCHIEL, 1997).

Para Okada e Ortega (2009, p. 19), a efetividade e a eficiência na recuperação de informações estão ligadas a tarefa de classificação, uma vez que "[...] não há habilidade de busca que supere modos inconsistentes de organização da informação".

Nas considerações de Martins e Carvalho (2014, p. 121), é incontestável que a qualidade da indexação pode interferir na eficácia do processo de busca nas bases de dados e que esta complexa tarefa é influenciada pelo conhecimento tácito e pela ontogenia do indexador (ou do arquiteto do *software*) "[...] para que o documento seja bem representado nas bases de dados.".

Em um entendimento inicial, recuperação da informação é o processo, manual ou automático (digital), pelo qual se retorna ao solicitante de uma pergunta formulada a informação armazenada que satisfaça sua necessidade informacional (LOPES, 2002).

O principal objetivo de um SRI é fornecer, rapidamente, aos usuários a informação que procuram. A complexidade em recuperar somente aqueles documentos que são importantes para o usuário é um dos principais obstáculos a ser contornado pelos SRI (FACHIN, 2009; RODRIGUES; CRIPPA, 2011; TEIXEIRA, 2010; WU *et al.*, 2013).

O crescimento exponencial do volume de informação levou a Recuperação da Informação (RI) estabelecer-se como um ramo do conhecimento científico que busca amenizar três problemas: representação da informação, especificação da busca da informação e criação de mecanismo para recuperação. Ressalta-se que a RI transita, de forma interdisciplinar, por diversos domínios, desde a Ciência da Informação

até Ciência da Computação, dispondo de tarefas e ferramentas de organização e recuperação da informação e conhecimento, como: classificação, tesauros, taxonomia e ontologias, entre outras (PONTES JUNIOR; CARVALHO; AZEVEDO, 2013).

No entendimento de Teixeira (2005, p. 81), este crescimento acelerado do volume de informações geradas e, por consequência, consumidas, conduziram ao emprego da TIC como uma das principais ferramentas no incremento da qualidade e da produtividade dos SRI, os quais devem "[...] atender às necessidades específicas dos usuários, permitindo ao máximo o acesso a informações relevantes".

Conforme Souza (2006[3] *apud* ARAUJO, 2012), as funções de um SRI são as seguintes: representar as informações contidas nos documentos e expressas pelos processos de indexação e descrição dos documentos; armazenar e gerir física e/ou logicamente esses documentos e suas representações; e, recuperar as informações ali contidas e os documentos armazenados no sistema.

Os SRI, sejam os utilizados para a pesquisa em escala *web* ou a um nível empresarial, impactam na vida diária e, no entanto, a intensidade e o sentido deste impacto são, raramente, medidos. Normalmente, sua avaliação é focada em seu motor de busca, medindo o quão rápido ele pode responder a uma consulta, ou o nível de relevância da informação recuperada.

As especificidades de busca de cada organização têm norteado a construção de mecanismos de recuperação inteligente de informação. Fachin (2009) e Strehl (2011) afirmam que a utilização de agentes inteligentes favorece o desenvolvimento de SRI que possam atender essas especificidades de acordo com o público alvo, e que cabe aos "[...] criadores investigar, analisar e utilizar estes recursos [...]" (FACHIN, 2009, p. 261).

Ainda neste sentido, Weikum *et al.* (2009) apontam para a convergência, tanto da perspectiva da recuperação da informação quanto da perspectiva dos bancos de dados, na utilização crescente de dados estruturados e semiestruturados.

A Web 2.0 e sua interatividade proporcionou aos usuários, através dos mais diversos aplicativos, contribuir para o aumento do volume de informações armazenadas, bem como, da quantidade de interessados na sua recuperação. Ghorab *et al.* (2012) observam que a maioria dos atuais SRI não considera as características do usuário que realiza a consulta; retornando o mesmo conjunto de

informações para usuários diferentes que realizem a mesma consulta. Entendem os referidos autores que estes sistemas deveriam considerar as características do usuário no desenvolvimento destes sistemas, os quais são denominados *Personal Information Retrieval* (PIR). Isso pode ser feito mantendo o controle de informações e interesses pessoais do usuário e, em seguida, usando essas informações para personalizar a consulta ou o conjunto de resultados apresentados.

O estudo realizado por Peltonen e Lin (2014) analisa o custo de uma tarefa de recuperação da informação baseada nos metadados, a fim de levantar o grau de similaridade entre os documentos recuperados; classificá-los; agrupá-los; e mostrá-los através de uma interface gráfica.

Nas considerações de Ramos e Munhoz (2011), um sistema de busca considerado "ideal" deverá dispor de funcionalidades que permitam a recuperação por partes de palavras (maq costu para máquina de costura, por exemplo), por sinônimos (chave de luz para interruptor, por exemplo), por parte alternada do todo e pela semântica da expressão fornecida pelo usuário.

Para Araujo, "[...] a função principal de um SRI é dispor de informações contidas nos documentos indexados, a partir de uma descrição sintética, objetiva e representativa de seu conteúdo formal e temático" (ARAUJO, 2012, p. 139-140). Teixeira e Schiel (1997) corroboram que estes sistemas integram vários processos, tais como: seleção, aquisição, indexação, busca e recuperação das informações.

As evoluções dos SRI, ocorridas nas mais diversas situações de tempo, espaço e necessidades organizacionais específicas, favoreceram o desenvolvimento de vários métodos de busca, dentre os quais destacamos o booleano, o espaço vetorial, o probabilístico, o *clustering* e o *feedback*. E para suprir as peculiaridades da Web Semântica, os modelos difuso, booleano estendido, espacial vetorial generalizado, indexação semântica latente, redes neurais e recuperação textual estruturada (FACHIN, 2009).

Neste trabalho, aborda-se, com alguma ênfase, o modelo espaço vetorial (ou simplesmente modelo vetorial) em virtude de seu emprego na concepção e desenvolvimento do motor de busca do MORE, possibilitando recuperar documentos que respondam, parcialmente, aos termos de uma busca.

Segundo os ensinamentos de Manning, Raghavan e Schütze (2009), o modelo espaço vetorial sugere um espaço no qual os termos de um documento e os termos da consulta são vetorizados de modo que o grau de similaridade entre eles seja calculado através de equações matemáticas e o conjunto de resultados ordenado, de acordo com este grau de similaridade.

Trata-se de um modelo estatístico e multidimensional, no qual cada termo representa uma dimensão e pode ter associado a ele um peso para refletir sua importância, tanto no documento quanto na consulta. Assim, o ângulo entre esses vetores determina o grau de similaridade entre eles, ou seja, o grau de similaridade é inversamente proporcional ao valor do ângulo formado entre eles.

Manning, Raghavan e Schütze (2009) elucidam, ainda, o fato de que documentos com os mesmos termos podem estar localizados em uma mesma região do espaço vetorial e, consequentemente, possuírem conteúdos similares.

Convém destacar que na fase de transformação dos dados, considerando as restrições de recursos computacionais e o objetivo principal da RI, esses são armazenados em um índice invertido que contém os termos e a frequência desses no documento e no *corpus*, para facilitar sua localização e recuperação. Para o caso em estudo, fez-se a carga inicial das referências existentes na base de dados e, a partir deste momento, cada referência inserida na base de dados terá também seus termos e frequência inseridos no índice invertido.

A fim de tirar proveito das características supracitadas, foi promovida a integração entre as ferramentas MORE e servidor de recuperação da informação *Apache Solr*®, cuja análise do impacto é objetivo do presente trabalho. Salienta-se que a análise proposta advém do ganho esperado na oferta do serviço aos usuários finais através da integração promovida entre o ferramental em tela.

3 Servidor de recuperação da informação *Apache Solr*®

A importância da recuperação da informação e seu emprego com foco na organização propiciou o desenvolvimento de várias soluções comerciais no âmbito da *Enterprise Search*, dentre as quais destacam-se: Google Search Appliance®,

Fast®, G2|B® e ISYS®, assim como, a opção de projetos *open source*, no caso do ElasticSearch® e do Solr®, este mantido pela Fundação Apache.

O Apache Solr® é um servidor de recuperação da informação para bases de dados textuais. É uma ferramenta desenvolvida em Java, *open source*, que faz uso da biblioteca de busca em texto completo conhecida como Lucene. Foi idealizado para ser uma aplicação *web* com a finalidade de prover várias funcionalidades de busca em texto completo, bem como, usar e ampliar as características da biblioteca *Apache Lucene*. Ambos, *Solr*® e *Lucene*®, são parte do projeto *Apache Lucene* e foram fundidos em um único esforço de desenvolvimento desde 2010, melhorando sua modularidade (KUMAR, 2013; SERAFINI, 2013).

Possibilita a exploração de suas funcionalidades através dos seus próprios serviços *REST-like*[4], que podem ser utilizados, de várias formas diferentes, e em, praticamente, qualquer plataforma ou linguagem. Há, ainda, a possibilidade de sua utilização como uma estrutura embarcada (*built-in*), acrescentando algumas de suas funcionalidades ao aplicativo Java em desenvolvimento, através de uma chamada direta à sua API interna.

Para melhorar o entendimento sobre como o Solr® implementa suas principais características, Serafini (2013) esclarece que é importante ter uma visão geral sobre o que é um índice Lucene (índice invertido), e como ele é feito. Conceitos fundamentais do Lucene são os seguintes:

a) documento - é a principal estrutura utilizada, tanto para as buscas como para os índices. Um documento é uma representação na memória dos valores dos dados que precisam ser utilizadas nas pesquisas. Para que isso funcione, todos os documentos consistem em um conjunto de campos, que é a estrutura de dados mais simples;

b) campo - tem seu próprio nome e valor, e é composto por, pelo menos, um termo. O campo é análogo ao atributo de uma relação em um Banco de Dados Relacional. Assim, cada documento pode ser visto como uma simples lista de pares campo-valor. Um campo pode ser monovalorado ou multivalorado. Se for monovalorado, ocorrerá a substituição do valor anterior à entrada de novo valor, enquanto que se for multivalorado, a entrada de novo valor será adicionado à lista de valores do campo;

c) termo - é a unidade básica para a indexação. Para simplificar, vamos imaginar uma única palavra, se o campo for do tipo texto, entretanto se for do tipo *string*, por exemplo, toda a sequência de caracteres (ou palavras) será indexada de maneira indivisível;

d) índice - é a estrutura em memória onde *Lucene*® e *Solr*® executam as buscas. Pode-se, então, pensar em um documento como um único registro no índice.

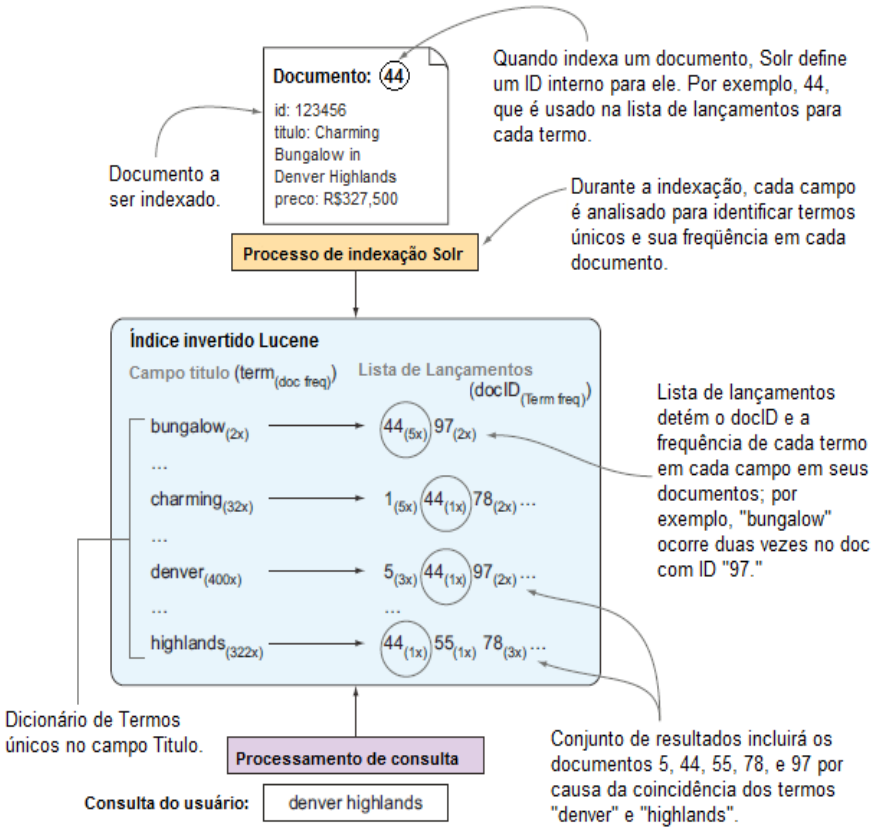
A Figura 1, na sequência, apresenta uma visão geral do processo de indexação.

Na argumentação de Grainger e Potter (2014, p. 4, tradução nossa), "*Solr*® é uma solução escalável, um mecanismo de busca organizacional pronto para entrar em produção, otimizado para busca de grandes volumes de dados centrado em texto e retornar resultados classificados por relevância". Está pronto para entrar em produção porque, além de ser *open source*, traz consigo um exemplo pré-configurado que auxilia e facilita sua instalação e configuração.

A escalabilidade do *software Solr*® fica bem evidenciada na sua versão 4 em que a funcionalidade *Cloud* permite a distribuição do trabalho (indexação e processamento da consulta) para vários servidores em um *cluster*.

Verifica-se que o *software Solr*® está otimizado para busca, uma vez que sua notável rapidez é observada ao executar consultas complexas em alguns centésimos de segundos. Está, também, apto a trabalhar com índices que contenham muitos milhões de documentos e sua otimização permite a busca em conteúdo expresso por meio de linguagem natural, tais como, mensagens eletrônicas, páginas da Web, currículos, documentos PDF e mensagens sociais - como *tweets* ou *blogs*.

Figura 1 - Estrutura de chave de dados para recuperação da informação.



Fonte: Adaptado de Grainger e Potter (2014).

No tocante à classificação dos resultados por relevância, verifica-se que o *Solr*® devolve documentos em ordem de classificação com base na relevância de cada documento, de acordo com a consulta do usuário. Faz uso de complexos cálculos matemáticos para a obtenção do escore de relevância, utilizando-se do conceito de Modelo de Espaço Vetorial. No entanto, a grande dificuldade centra-se na atribuição do valor (peso) a cada termo e/ou campo utilizado na consulta. Tal fato exige conhecimentos especializados do usuário na elaboração de sua consulta. Todavia, isto pode ser amenizado com a utilização de interfaces gráficas que possibilitem, por meio de listas pré-formatadas, que o usuário expresse suas considerações.

Atualmente, o *software Solr*® encontra-se em sua versão de número 6 e a cada versão adiciona novas funcionalidades que satisfazem as necessidades atuais requeridas, para que os SRI atendam aos requisitos organizacionais. Pode-se afirmar que se trata de uma tecnologia madura e flexível, que oferece, além de complexas características de busca em texto completo, as funcionalidades de

autossugestão, filtragem avançada, pesquisa geocodificada, destaque em texto, pesquisa facetada, entre muitas outras.

Além disso, está sendo utilizado por diversas organizações nos mais variados cenários. A título exemplificativo cita-se o *The Guardian*®, grandes empresas (Apple®, Disney®, Goldman Sachs®), *sites* de notícias, *sites* institucionais (como *sites* do governo Americano), *sites* de editoras, entre outros. Percebe-se que nem todos os *sites* são de linguagem natural em texto completo, o que nos faz crer que as outras características do *Apache Solr*® influenciaram na escolha (SERAFINI, 2013).

Destaca-se, novamente, a possibilidade de integração do *Apache Solr*® com outras aplicações, utilizando-se as mais variadas linguagens de programação. Listam-se, a seguir, os três principais projetos desenvolvidos em PHP, uma vez que o MORE faz uso dessa linguagem de programação e porque são *open source*. São eles: *Solr-PHP-client*, *Apache Solr-PHP extension* e *Solarium*.

O *Solr-PHP-client* é um projeto descontinuado, uma vez que sua última atualização ocorreu em maio de 2011. É um cliente com funcionalidades básicas e que não suporta vários recursos implementados nas versões mais recentes do *Apache Solr*®.

O projeto *Apache Solr-PHP extension*, por sua vez, trata-se de uma biblioteca leve, rápida e dispõe de variados recursos que possibilitam a comunicação entre aplicações desenvolvidas em PHP e instâncias do servidor *Apache Solr*® por meio de uma API (*Application Programming Interface*) orientada a objetos.

Kumar (2013, p. 14, tradução nossa) foi enfático ao afirmar que "*Solarium* é a mais recente biblioteca para integração *Solr*®/PHP, além de ser atualizada continuamente.". Observa-se que a afirmação acima continua válida atualmente, uma vez que não foi possível identificar nenhum projeto mais recente com a proposição de promover a integração entre aplicações desenvolvidas em PHP e o *Apache Solr*®. Foi desenvolvida no paradigma de programação orientada a objetos e dispõe de recursos atualizados em consonância com o *Apache Solr*®. Sua flexibilidade é evidenciada ao permitir a adição de uma funcionalidade que se faça necessária à aplicação. Também permite parametrização personalizada para atingir quase todas as

tarefas, aliviando o usuário na lide com todos os complexos parâmetros de consulta do *Apache Solr*®, além de dispor de uma vasta documentação.

Das alternativas apresentadas, optou-se pelo *Solarium* para promover o suporte ao desenvolvimento do protótipo utilizado nessa pesquisa. A escolha ocorreu em função da curva de aprendizado, da usabilidade, das funcionalidades disponíveis e das atualizações do *Solarium* que procuram dar suporte a todos os novos recursos disponíveis no *Apache Solr*®.

4 Procedimentos metodológicos

Nesta seção, serão abordados as técnicas e os métodos utilizados na realização da pesquisa.

A pesquisa foi desenvolvida segundo os preceitos da Metodologia da Experimentação, que para Cervo, Bervian e Silva (2010, p. 39) "[...] é o conjunto de processos utilizados para verificar as hipóteses". Esclarecem, ainda, os autores que "[...] o princípio geral no qual se fundamentam as técnicas da experimentação é o do determinismo, que se anuncia assim: nas mesmas circunstâncias, as mesmas causas produzem os mesmos efeitos". A aplicação dessa metodologia materializou-se por meio da realização de experimentos realizados nos laboratórios da instituição, com o objetivo de verificar a regularidade, ou não, dos resultados.

Inicialmente, foi construído um referencial teórico, por meio de uma pesquisa bibliográfica, a fim de situar o objeto da pesquisa em relação ao estado da arte. Severino (2007) ensina que a pesquisa bibliográfica é a realizada em registros de trabalhos anteriores, cujas contribuições encontram-se disponibilizadas em relatórios impressos e/ou digitais, tais como: livros, artigos e teses, entre outras fontes bibliográficas.

No passo seguinte, foram realizadas as tarefas referentes à instalação e à configuração customizada do SRI *Apache Solr*® em um servidor de aplicativos *Java Apache Tomcat*® 8. Salienta-se que a configuração do *Solr*® é realizada através de uma coleção de arquivos do tipo XML, do inglês *eXtensible Markup Language*, que é uma linguagem de marcação recomendada pelo W3C[5] e utilizada na criação de documentos com dados organizados, seguindo uma determinada hierarquia.

A fase posterior constou da recuperação dos dados armazenados na relação 'referencia'[6] do banco de dados relacional MySQL do MORE através de um *script* em PHP que permitiu percorrer os registros e gerar arquivos tipo JSON (*JavaScript Object Notation*) com uma centena de documentos cada, para posterior indexação. Assim, cada registro da relação é transformado em um documento a ser indexado. Este documento consta de três campos: 'CHAVEREFERENCIA' (*unique key*), 'REFE' (campo de texto que contém uma referência bibliográfica gerada) e 'TIPOTABELA' (identifica a relação que armazena os dados informados pelo usuário para a geração de uma referência bibliográfica).

Salienta-se que, apesar do espaço ocupado em disco (ou em memória) pela tabela 'referência' não ultrapassar 4 GB, o tempo demandado para a resposta a uma busca é elevado, pelo fato do gerenciador de banco de dados percorrer todos os registros e compará-los com os termos da consulta, o que não ocorre com o *Apache Solr*®, como pode ser observado na Figura 1, acima.

Nas considerações de Severance (2012, p. 6, tradução nossa), "As estruturas mais comuns que usamos na programação são variáveis escalares, listas lineares, e pares de chave-valor.". JSON representa essas estruturas na serialização, natural e diretamente, reduzindo o consumo de recursos computacionais entre as estruturas em memória em aplicativos e o formato de serialização. Uma implementação JSON em *JavaScript* lhe dá uma vantagem distinta sobre outros formatos de serialização, como XML, quando se trabalha com aplicações parcialmente escritas em *JavaScript*. Pode-se afirmar que o sucesso de JSON dá-se ao fato de que as estruturas de dados, por ele representadas, são exatamente as mesmas estruturas de dados que as linguagens de programação representam.

Utilizou-se o formato JSON pelos motivos acima e pela disponibilidade de funções nativas do PHP para manipular essas estruturas.

Na etapa seguinte, ocorreu a inserção dos documentos na base de dados do *Apache Solr*®, a fim de que o *Apache Lucene*® pudesse construir o índice invertido, adicionando-os e indexando-os para posterior utilização na recuperação da informação. Essa tarefa foi realizada com o apoio do *software cURL*, que provê uma biblioteca e uma ferramenta de linha de comando para transferência de dados, com suporte a vários protocolos, inclusive o HTTP[7], que foi o utilizado.

No quinto passo, foi desenvolvido o protótipo com a finalidade de agilizar o processo de busca e disponibilização do resultado ao interessado. Na construção do protótipo, foi utilizada a biblioteca *Solarium*, facilitando a comunicação e a integração entre o *Apache Solr*® e o protótipo. Vale salientar que, optando-se pela utilização desse novo mecanismo de recuperação da informação, as tarefas de indexação de inserção de novos documentos, ou mesmo, atualização ou exclusão, no índice invertido já estão desenvolvidas, faltando apenas a integração com o MORE.

Nas fases subsequentes, coletaram-se os dados, e fez-se uma análise ponderada, de modo que fique esclarecido de que forma os dados coletados contribuíram para o cumprimento do objetivo geral da pesquisa.

5 Resultados e discussões

O processo de avaliação ocorreu por meio da medição do tempo de resposta às consultas realizadas, tanto através do sistema atual como através do protótipo que faz uso das funcionalidades do *Apache Solr*®. A credibilidade da medição fundamenta-se no fato dessa ter sido implementada via código de programação em ambos ambientes considerados, ou seja, armazena-se o instante imediatamente anterior ao início da tarefa e o instante imediatamente posterior ao cumprimento da tarefa. A partir disto, é obtida a diferença, e ela é apresentada juntamente com o conjunto de informações recuperadas. Sendo assim, considerou-se a precisão de milésimo de segundo suficiente para a pesquisa em andamento.

Nesse sentido, foram elaboradas e submetidas 30 consultas distintas, as quais foram repetidas, com a finalidade de verificar a efetividade do *cache* de consultas realizadas no *Apache Solr*®. O número de documentos armazenados em *cache* é definido através do arquivo de configuração 'solrconfig.xml' para cada instância do *Apache Solr*®, impactando, diferentemente, em cada caso no consumo de recursos computacionais do servidor hospedeiro. No entanto, o que se verifica é que o espaço de memória utilizado nesta funcionalidade é relativamente pequeno. No caso em estudo, manteve-se a configuração padrão de 512 consultas em *cachê*, e considerando que cada consulta ocupa 5 KB de espaço, o impacto seria o ocasionado pelo consumo de 2,5 MB. Os dados colhidos

encontram-se nas Tabelas 1 e 2, discutidas a seguir. Considerou-se a quantidade de 30 consultas o suficiente para que se tenha uma tendência de tempos de resposta, aliado à restrição de espaço para apresentação dos resultados obtidos.

A recuperação da informação, no sistema atual, é feita através de consulta SQL à base de dados MySQL, e a sequência de caracteres (palavras) informada pelo usuário é submetida através da restrição 'LIKE'[8], o que significa a exigência de coincidência exata entre a sequência submetida e alguma outra idêntica que possa ser localizada no atributo 'REFE' da relação 'referência' da base de dados MySQL. O conjunto de resultados é ordenado de acordo com a sua posição na relação de origem, ou seja, não considera o grau de relevância.

A recuperação da informação, através do protótipo, utiliza-se das características suportadas e disponibilizadas pelo *Apache Solr*®, cuja mais evidente é a busca rápida em texto completo. Para organizar o conjunto de resultados segundo o grau de relevância, foi lançado mão da funcionalidade 'edismax' na elaboração da consulta, atribuindo peso[9] 6 aos termos encontrados no campo 'CHAVEREFERENCIA' e peso 1,5 aos termos encontrados no campo 'REFE'. Além de configurar a recuperação apenas dos documentos que contenham, no mínimo, 50% dos termos informados.

Tabela 1 - Tempo de Resposta (TR) à submissão dos Termos de Consulta (Consulta Inicial).

Nr Ord.	Termos da Consulta	Sistema Atual		Protótipo	
		TR em s.	Qtd. Reg.	TR em s.	Qtd. Reg.
01	Automação Industrial	19,906	+ de 100	0,955	+ de 100
02	Tecnologias da Informação	17,459	+ de 100	1,850	+ de 100
03	Eletrônica Digital	16,840	+ de 100	0,771	+ de 100
04	Recuperação da Informação	18,788	+ de 100	0,769	+ de 100
05	Web Semântica	17,864	+ de 100	0,898	+ de 100
06	Representação do Conhecimento	18,159	+ de 100	1,472	+ de 100
07	Semicondutores	18,183	+ de 100	0,316	+ de 100
08	Metodologia do Trabalho Científico	17,193	+ de 100	1,190	+ de 100
09	Modelo Vetorial	16,376	1	0,799	+ de 100
10	Mineração de Dados	18,218	+ de 100	0,866	+ de 100
11	Estrutura de Dados	17,522	+ de 100	0,794	+ de 100
12	Sistemas Interativos	17,518	28	0,696	+ de 100
13	Prentice Hall	17,703	+ de 100	1,072	+ de 100
14	Engenharia de Software	17,595	+ de 100	1,235	+ de 100
15	Sistema Operacionais Embarcados	17,175	1	0,749	+ de 100
16	Programação em Computadores	17,088	1	1,735	+ de 100
17	Redes de Computadores	17,463	+ de 100	0,575	+ de 100
18	Ivan Sommerville	15,856	0	0,353	+ de 100
19	Maurício Samy Silva	15,504	8	0,797	+ de 100
20	Juliano Niederauer	15,802	5	0,708	+ de 100
21	Web Service PHP	17,920	0	0,737	+ de 100
22	Mundo Virtual 3D	17,382	3	0,939	+ de 100
23	Big Data	17,562	+ de 100	0,670	+ de 100
24	Big Table	15,362	0	0,316	+ de 100
25	Redes Neurais Artificiais	17,478	+ de 100	0,560	+ de 100
26	Mecanismo Online para Referências	18,670	+ de 100	0,958	+ de 100
27	Fundamentos Filosóficos da Pesquisa	15,860	0	1,051	+ de 100
28	Ambientes Virtuais de Aprendizagem	17,314	+ de 100	0,753	+ de 100
29	Inteligência Artificial	20,833	+ de 100	0,265	+ de 100
30	Algoritmo Genético	16,899	99	0,473	+ de 100

Fonte: elaborada pelos autores.

No tocante à quantidade de documentos/registros recuperados, em ambos os casos foi estabelecido um limite de 100 ocorrências. Esse limite deve-se ao fato de que os resultados apresentados aos usuários são paginados de 20 em 20 registros, totalizando cinco páginas; e que segundo iProspect (2006), em seu trabalho, afirma que mais de 90% dos usuários não vão além da terceira página de resultados a busca de informações por eles realizada.

Nos dados tabulados na Tabela 1, verifica-se que o tempo de resposta médio no sistema atual é de 17,450 segundos e que no protótipo é de 0,844

segundos, estabelecendo uma relação de tempo entre ambos de, aproximadamente, 20 vezes, ou seja, o tempo de resposta do sistema atual é 20 vezes maior que o tempo de resposta do protótipo. O protótipo recuperou mais de 100 documentos para todas as consultas realizadas e as ordenou de acordo com o grau de relevância. O mesmo não aconteceu com o sistema atual, retornando, em alguns casos, um conjunto vazio pelos motivos especificados na sequência.

Tabela 2 - Tempo de Resposta (TR) à submissão dos Termos de Consulta (Repetição da Consulta).

Nr Ord.	Termos da Consulta	Sistema Atual		Protótipo	
		TR	Qtd. Reg.	TR	Qtd. Reg.
01	Automação Industrial	17,296 s	+ de 100	0,060 s	+ de 100
02	Tecnologias da Informação	16,669 s	+ de 100	0,080 s	+ de 100
03	Eletrônica Digital	17,472 s	+ de 100	0,057 s	+ de 100
04	Recuperação da Informação	17,986 s	+ de 100	0,055 s	+ de 100
05	Web Semântica	18,042 s	+ de 100	0,067 s	+ de 100
06	Representação do Conhecimento	17,668 s	+ de 100	0,054 s	+ de 100
07	Semicondutores	18,965 s	+ de 100	0,057 s	+ de 100
08	Metodologia do Trabalho Científico	17,529 s	+ de 100	0,056 s	+ de 100
09	Modelo Vetorial	16,815 s	1	0,068 s	+ de 100
10	Mineração de Dados	17,279 s	+ de 100	0,058 s	+ de 100
11	Estrutura de Dados	18,132 s	+ de 100	0,078 s	+ de 100
12	Sistemas Interativos	17,873 s	28	0,061 s	+ de 100
13	Prentice Hall	16,686 s	+ de 100	0,058 s	+ de 100
14	Engenharia de Software	18,661 s	+ de 100	0,053 s	+ de 100
15	Sistema Operacionais Embarcados	16,431 s	1	0,057 s	+ de 100
16	Programação em Computadores	15,822 s	1	0,071 s	+ de 100
17	Redes de Computadores	16,608 s	+ de 100	0,057 s	+ de 100
18	Ivan Sommerville	17,860 s	0	0,057 s	+ de 100
19	Maurício Samy Silva	16,022 s	8	0,068 s	+ de 100
20	Juliano Niederauer	16,475 s	5	0,050 s	+ de 100
21	Web Service PHP	17,390 s	0	0,059 s	+ de 100
22	Mundo Virtual 3D	16,389 s	3	0,054 s	+ de 100
23	Big Data	16,501 s	+ de 100	0,056 s	+ de 100
24	Big Table	16,170 s	0	0,053 s	+ de 100
25	Redes Neurais Artificiais	16,618 s	+ de 100	0,062 s	+ de 100
26	Mecanismo Online para Referências	18,507 s	+ de 100	0,054 s	+ de 100
27	Fundamentos Filosóficos da Pesquisa	17,230 s	0	0,054 s	+ de 100
28	Ambientes Virtuais de Aprendizagem	17,451 s	+ de 100	0,055 s	+ de 100
29	Inteligência Artificial	17,625 s	+ de 100	0,056 s	+ de 100
30	Algoritmo Genético	16,993 s	99	0,054 s	+ de 100

Fonte: elaborada pelos autores.

Para a consulta cujos termos são ‘Ivan Sommerville’, observa-se que o sistema atual retornou um conjunto vazio em virtude das normas da ABNT determinarem que o sobrenome do autor deve anteceder o nome, não permitindo nenhuma ocorrência idêntica à consulta submetida. Para os outros dois autores, aconteceram algumas ocorrências em virtude de constarem como tradutor ou como organizador, esta de maneira equivocada.

No caso de 'Web Service PHP', verifica-se que o real nome do livro é 'Web Service em PHP', situação que retornaria um conjunto não vazio.

Nos dados tabulados na Tabela 2, verifica-se que o tempo de resposta médio no sistema atual é de 16,739 segundos e que no protótipo é de 0,059 segundos, estabelecendo uma relação de tempo entre ambos de, aproximadamente, 280 vezes; ou seja, o tempo de resposta, do sistema atual, é 280 vezes maior que o tempo de resposta do protótipo. Observa-se que a característica do *Apache Solr*®, e explorada pelo protótipo, de manter em *cache* o resultado das últimas consultas realizadas agiliza a recuperação da informação, melhorando o tempo de resposta. No presente caso, a parcela de tempo utilizada na repetição da consulta foi de apenas 1/14 em relação à primeira consulta, ou seja, o tempo de resposta na repetição de uma consulta anterior é de apenas 1/14 do tempo de resposta da primeira consulta. No caso do sistema atual, que não se utiliza de *cache* de consultas realizadas, na repetição da consulta, os tempos de resposta oscilaram em torno dos tempos observados na primeira consulta.

6 Considerações finais

O aumento do volume de documentos gerados e armazenados a partir do pós-guerra e ampliado com o advento da Web 2.0, aliado à necessidade diária de tomada de decisão no âmbito organizacional, propiciaram o desenvolvimento da RI. Para dar suporte a esta rotina diária foram concebidos e são, constantemente atualizados, os SRI.

A sobrecarga informacional digital verificada na atualidade advém, não somente da produção de recursos informacionais originalmente digitais, como também de processos de digitação e/ou digitalização de recursos produzidos em outras formas de armazenamento da informação, como, por exemplo, o impresso em papel.

A importância destes sistemas no cotidiano organizacional contribuiu para o surgimento de várias soluções comerciais e alguns projetos *open source*. Dentre esses projetos *open source* destaca-se o *Apache Solr*®, que se utiliza do índice invertido viabilizado pelo *Apache Lucene*® e da coincidência parcial dos termos da

consulta na recuperação da informação em bases textuais, o qual foi utilizado na construção do protótipo desenvolvido para servir aos propósitos desta pesquisa.

Da análise dos dados constantes das Tabelas 1 e 2 da Seção 5 deste documento, verifica-se que a utilização do servidor de recuperação da informação *Apache Solr*® impactou, positivamente, no tempo de resposta das consultas realizadas e foi ainda melhor na repetição destas consultas, melhorando, significativamente, o tempo de resposta às consultas submetidas ao motor de busca do MORE.

Verifica-se que o modelo espaço vetorial, utilizado pelo *Apache Solr*® para atribuir um grau de relevância para cada documento recuperado, permite a recuperação por coincidência parcial, e a ordenação do conjunto ocorre segundo o grau de relevância. Assim sendo, mesmo que o usuário cometa algum equívoco na digitação dos termos que compõem a consulta, o sistema calcula o grau de relevância em função dos termos corretos, evitando frustrações indesejadas.

Da relação de tempo de resposta na recuperação da informação entre o sistema atual e o protótipo proposto depreende-se, para o caso desta pesquisa, que a utilização de recursos de TIC que satisfazem o conceito de *Enterprise Search* impacta, positivamente, no processo de recuperação da informação contida na base de dados do MORE.

Das considerações feitas ao longo deste trabalho e da análise dos dados coletados, foi possível verificar que a utilização de motores de busca, baseados no *Apache Solr*®, impacta, positivamente, no processo de recuperação da informação contida na base de dados do MORE.

Os resultados obtidos durante a pesquisa serviram de suporte à tomada de decisão de integrar o protótipo ao sistema atual. Salienta-se, ainda, que o mesmo já se encontra em produção em http://more.ufsc.br.

Ao finalizar esse trabalho identifica-se a possibilidade de melhoria na usabilidade e na ergonomia do MORE por meio da implementação da funcionalidade de autossugestão viabilizada pelo *Apache Solr*®.

Referências

ALVES, Maria Bernadete Martins; MENDES, Leandro Luiz; ALVES, João Bosco da Mota. MORE: Mecanismo On-line para Referências. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 14., 2006, Salvador. **Anais**... Salvador: UFBA, SIBI, 2006. p. 1-12. Disponível em: <http://xa.yimg.com/kq/groups/25169972/1250144979/name/4.ALVES,M.+B.+M.pdf>. Acesso em: 12 mar. 2016.

ARAUJO, Vera Maria Araujo Pigozzi de. Sistemas de recuperação da informação: uma discussão a partir de parâmetros enunciativos. **Transinformação**, Campinas, v. 2, n. 24, p. 137-143, 2012.

CARVALHO, Lidiane dos Santos; LUCAS, Elaine R. de Oliveira; GONÇALVES, Lucas Henrique. Organização da informação para recuperação em redes de produção e colaboração na Web. **Revista ACB**: Biblioteconomia em Santa Catarina, Florianópolis, v. 15, n. 1, p. 71-86, jun. 2010.

CERVO, Amado Luiz; BERVIAN, Pedro Alcino; SILVA, Roberto da. **Metodologia científica**. 6. ed. São Paulo: Pearson, 2010.

FACHIN, Gleisy Regina Bories. Recuperação inteligente da informação e ontologias: um levantamento na área da Ciência da Informação. **Biblos**: revista do Instituto de Ciências Humanas e da Informação, Rio Grande, v.23, n.1, p. 259-283, 2009.

GHORAB, M. Rami et al. Personalised information retrieval: survey and classification. **User Modeling and User-Adapted Interaction**, Dordrecht, v. 23, n. 4, p. 381-443, May 2012.

GRAINGER, Trey; POTTER, Timothy. **Solr in action**. Shelter Island: Manning, 2014.

IPROSPECT. **iProspetc search engine user behavior study**. [S.l.], 2006. Disponível em: <http://district4.extension.ifas.ufl.edu/Tech/TechPubs/WhitePaper_2006_SearchEngineUserBehavior.pdf>. Acesso em: 15 nov. 2015.

KUMAR, Jayant. **Apache Solr PHP integration**. Birmingham: Packt, 2013.

LOPES, Ilza Leite. Estratégia de busca na recuperação da informação: revisão da literatura. **Ciência da Informação**, Brasília, DF, v. 31, n. 2, p. 60-71, ago. 2002.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **An introduction to information retrieval**. Draft: Cambridge University Press,

2009. Disponível em: <http://www-nlp.stanford.edu/IR-book/>. Acesso em: 22 dez. 2014.

MARTINS, Elaine Cristina Domingues; CARVALHO, Tatiana. Recuperação da informação em psicologia: LILACS e Index Psi Revistas Técnico-Científicas. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 19, n. 2, p.118-130, jun. 2014.

OKADA, Susana Yuri; ORTEGA, Cristina Dotta. Análise da recuperação da informação em catálogo online de biblioteca universitária. **Informação & Informação**, Londrina, v. 14, n. 1, p.18-35, ago. 2009.

OLIVEIRA, Dalgiza Andrade; ARAUJO, Ronaldo Ferreira de. Construção de linguagens documentárias em sistemas de recuperação da informação: a importância da garantia do usuário. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 17, n. 34, p.17-30, ago. 2012.

PELTONEN, Jaakko; LIN, Ziyuan. Information retrieval approach to meta-visualization. **Machine Learning**, Boston, v. 99, n. 2, p. 189-229, Oct. 2014.

PONTES JUNIOR, João de; CARVALHO, Rodrigo Aquino de; AZEVEDO, Alexander William. Da recuperação da informação à recuperação do conhecimento: reflexões e propostas. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 18, n. 4, p.2-17, dez. 2013.

RAMOS, Clériston; MUNHOZ, Deise Parula. A subjetividade da relevância na recuperação da informação: análise a partir de imagens representativas. **Biblos**: revista do Instituto de Ciências Humanas e da Informação, Rio Grande, v. 25, n. 1, p.69-79, jun. 2011.

RIBEIRO, Fernanda. O uso da classificação nos arquivos como instrumento de organização, representação e recuperação da informação. In: CONGRESSO ISKO ESPANHA E PORTUGAL, 1., 2013, Porto. **Informação e/ou conhecimento**: as duas faces de Jano. Porto: Cetac Media, 2013. p. 528-539.

RODRIGUES, Bruno César; CRIPPA, Giulia. A recuperação da informação e o conceito de informação: o que é relevante em mediação cultural? **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 16, n. 1, p. 45-64, 2011.

SERAFINI, Alfredo. **Apache Solr beginner's guide**. Birmingham: Packt, 2013.

SEVERANCE, Charles. Discovering javascript object notation. **Computer**, Long Beach, v. 45, n. 4, p. 6-8, Apr. 2012.

SEVERINO, Antônio Joaquim. **Metodologia do trabalho científico**. 23. ed. São Paulo: Cortez, 2007.

STREHL, Leticia. As folksonomias entre os conceitos e os pontos de acesso: as funções de descritores, citações e marcadores nos sistemas de recuperação da informação. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 16, n. 2, p. 101-114, 2011.

TEIXEIRA, Cenidalva Miranda de Sousa; SCHIEL, Ulrich. A internet e seu impacto nos processos de recuperação da informação. **Ciência da Informação**, Brasília, v. 26, n. 1, p. 9-20, 1997.

TEIXEIRA, Fábio Augusto Guimarães. **A recuperação da informação e a colaboração de usuários na Web**. 2010. 160 f. Dissertação (Mestrado)-Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, 2010.

TEIXEIRA, Robson da Silva. Serviço de recuperação da informação na biblioteca de um laboratório farmacêutico: um estudo prático. **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 2, n. 2, p. 80-89, jun. 2005.

WEIKUM, Gerhard et al. Database and information- retrieval methods for knowledge discovery. **Communications of the ACM**, New York, v. 52, n. 4, p. 56-64, Apr. 2009.

WHITE, Martin. **Enterprise search**: enhancing business performance. Sebastopol: O'Reilly Media, 2013.

WU, Mingfang et al. Cost and benefit estimation of experts' mediation in an enterprise search. **Journal of the Association for Information Science and Technology**, New York, v. 65, n. 1, p.146-163, Oct. 2013.

XAVIER, Raphael Figueiredo. **Análise de métodos de produção de interfaces visuais para recuperação da informação**. 2009. 78 f. Dissertação (Mestrado)-Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2009.

Information retrieval and database query in the search process of the Online Mechanism for References

Abstract: The data of organizations grows exponentially each year and has brought increment to the administrators and managers in the decision-making process, which they are daily submitted. In order to manage this data and discover information contained on them, the Information Retrieval Systems is widely used

in the organizational environment. The Information Retrieval was mainly developed to provide quickly to the users the information they seek. The evaluation of an Information Retrieval System is focused on its search engine, measuring how fast it can respond to a query, or the relevance level of the retrieved information. This study verifies the impact of using search engines, based on Apache Solr®, at the information retrieval process contained in the Mecanismo Online para Referências database. Thus, we researched the literature, searching for fundamentals to conceptualize the Information Retrieval and to deal with the peculiarities that are consistent with the scope of this research. We analyze the main features of the Apache Solr® Information Retrieval Server and the developed prototype built to evaluate this study. It should be clarify that the Apache Solr® was set up to sort the results by relevance level, and the Vector Space Model was used to calculate the degree of similarity. After that, the collected data is tabulated, presented and analyzed. We conclude that the use of search engines, based on Apache Solr®, impacts positively on the information retrieval process contained in the Mecanismo Online para Referências database.

Keywords: Information Retrieval. Indexing. Vector Space Model. Mecanismo Online para Referências.

Recebido: 09/04/2016

Aceito: 21/06/2016

[1] Web 2.0 é o termo utilizado para referir-se à segunda geração da *World Wide Web* (WWW) e às suas funcionalidades, tais como: interatividade, interoperabilidade e dinamicidade, entre outras. Possibilitou o surgimento dos ambientes colaborativos e interativos a exemplo das *wikis*, das redes sociais e dos blogs.

[2] Considera-se que *Enterprise Search* é um aplicativo de pesquisa empresarial que permite encontrar todas as informações que a empresa possui, sem a necessidade de saber onde estão armazenadas (WHITE, 2013).

[3] SOUZA, Renato Rocha. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, Belo Horizonte, v.11, n.2, p.161-173, maio/ago. 2006.

[4] O termo *REST-like* refere-se aos serviços que aderem à arquitetura REST, no entanto, não satisfazem todos os seus princípios.

[5] O W3C (*World Wide Web Consortium*) é uma comunidade internacional, onde as organizações associadas, uma equipe em tempo integral e o público interessado trabalham em conjunto para

desenvolver padrões *web*. A missão do W3C é conduzir a Web ao seu potencial máximo. Disponível em: <https://www.w3.org/Consortium/>. Acesso em: 13 jun. 2016.

[6] A relação 'referencia' é a tabela da base de dados, cujo nome é 'referencia', onde são armazenadas as referências bibliográficas após serem processadas. Os dados informados pelos usuários são armazenados em outras tabelas, de acordo com a fonte bibliográfica.

[7] HTTP (*Hypertext Transfer Protocol*) é um protocolo de comunicação utilizado entre sistemas distribuídos de hipermídia, que possibilita a leitura não linear através de enlaces que conduzem à apresentação dos recursos informacionais a eles vinculados. Cita-se como exemplo desse tipo de sistema a *World Wide Web*.

[8] A restrição 'LIKE' permite que uma consulta feita ao banco de dados retorne os registros que contenham exatamente a sequência de caracteres informada. Por exemplo: 'SELECT * FROM referencia WHERE REFE LIKE '%mecanismo%'', retornaria todos os registros da tabela 'referencia' cuja palavra 'mecanismo' fosse encontrada na coluna 'REFE'. O símbolo '%', neste caso, permite a ocorrência de outros caracteres antes e/ou depois da palavra 'mecanismo'.

[9] O peso dos termos de busca, neste trabalho, foi atribuído empiricamente. A atribuição de peso 6 aos termos encontrados no campo 'CHAVEREFERENCIA' baseia-se no fato de este ser um campo de chave única.