

Uma Interface de Programação de Aplicativos para análise qualitativa de textos aplicada ao sistema MineraFórum

Lucas Baruffi¹, Sílvio César Cazella², Sandro José Rigo¹, Jorge L. V. Barbosa¹

¹UNISINOS – Universidade do Vale do Rio dos Sinos

²UFCSPA – Universidade Federal de Ciências da Saúde de Porto Alegre

lucas.baruffi@gmail.com, silvioc.ufcspa@gmail.com, rigo@unisinobr, jbarbosa@unisinobr

Abstract. *This paper presents study of MineraFórum System and evaluation of its application in the integration of an Application Programming Interface (API) to text mining. A software reengineering was performed in order to extract the algorithm and construct the API. An system was constructed, in Java programming language, to evaluate the results, that show relevant aspects in its use with several documents, larger that the discussion fórum messages, original objective of the algorithm.*

Resumo: *Este artigo tem como objetivo apresentar o estudo do algoritmo de mineração de textos utilizado no sistema MineraFórum e avaliar a possibilidade de sua utilização para a construção de uma biblioteca de mineração de textos. A motivação principal está associada com a ampliação de utilização deste algoritmo. Foi realizada uma reengenharia de software no código do sistema para extrair a parte do seu algoritmo relevante à mineração de qualquer tipo de texto e criado um aplicativo para testar os resultados gerados. A implementação utilizou a linguagem Java, já utilizada pelo MineraFórum. Os resultados obtidos são relevantes e pôde-se comprovar que o algoritmo do MineraFórum tem potencial para avaliar textos maiores e não somente postagens em fóruns de discussão, foco original do algoritmo MineraFórum.*

1. Introdução

A WEB constitui-se em um dos meios de compartilhamento de informação mais difundido na atualidade. Como boa parte da informação disponibilizada na WEB não se apresenta estruturada, por vezes é necessária a leitura de documentos para a identificação daqueles conteúdos que são do interesse do usuário. As técnicas de Mineração de Textos atuam de modo a extrair conhecimento novo a partir de grandes quantidades de documentos não estruturados (Oliveira, 2012; Silva et al, 2012). Atividades ligada à educação podem ser beneficiadas com ferramentas que apoiem a identificação de aspectos em documentos em formato textual (Rigo et al, 2014).

O sistema MineraFórum (Azevedo, 2011) trata de identificar a qualidade das mensagens textuais em fóruns de discussão em ambientes virtuais de ensino e aprendizagem. O objetivo deste sistema é auxiliar os professores a avaliarem a relevância da postagem de mensagens textuais de alunos em relação ao tema proposto para uma discussão, no contexto de uma disciplina. O sistema apresenta como resultado

da análise das postagens feitas pelos alunos, um grafo relacionando os termos de maior relevância frente ao tema proposto e também gera um grafo com a postagem do aluno. Através de um cálculo de proximidade que verifica a relação entre o que foi postado pelo aluno e o que foi proposto no tema, é possível verificar a qualidade das postagens nesse fórum. Os resultados obtidos pelo MineraFórum em testes foram muito satisfatórios, conforme descreve artigo de Azevedo et al. (2011). O sistema foi implementado em forma de plugin para plataformas específicas de Educação à Distância como, por exemplo, o Moodle.

Tendo em vista os bons resultados obtidos na análise qualitativa das postagens, surgiu a necessidade de converter o plugin em um sistema autônomo que pudesse fazer essa análise para qualquer fonte de conteúdo e não apenas de postagens em fóruns de Ambientes Virtuais de Ensino e Aprendizagem, desta forma colaborando em necessidades de localização de materiais adequados para estudo, o que pode ser difícil atualmente, em função do grande volume de documentos existente. Como o código fonte do sistema está disponível para estudo, foi possível utilizar os algoritmos já implementados e testados para integrar uma Interface de Programação de Aplicativos (API) para análise de conteúdo. O código fonte do MineraFórum foi escrito na linguagem JAVA e a solução proposta para esse trabalho também foi desenvolvida nessa linguagem.

O trabalho tomou como princípios aqueles indicados por Thomas Scheller e Eva Kuhn (2012), que realizam um estudo minucioso sobre usabilidade de bibliotecas de programação, mostrando como diferentes conceitos de arquitetura influenciam na eficiência e satisfação dos programadores. O resultado desse trabalho nos mostra as dificuldades encontradas pelos programadores ao utilizar diferentes arquiteturas, bem como sugestões na organização de classes e métodos.

O objetivo deste artigo é apresentar o desenvolvimento de uma Interface de Programação de Aplicativos (API) que permita a criação de ferramentas de análise qualitativa de conteúdos em textos. O modelo utilizado nesta biblioteca de programação foi desenvolvido e validado por Azevedo (2011), originalmente destinado a verificar a relevância de mensagens curtas em fóruns de discussão no contexto de Ambientes Virtuais de Ensino e Aprendizagem. Para tal, foi necessário realizar um estudo aprofundado do algoritmo utilizado originalmente, denominado MineraFórum, para em seguida desenvolver uma API capaz de ser integrada em outros aplicativos. A avaliação do trabalho desenvolvido envolveu a criação de um protótipo capaz de realizar a análise de documentos de texto a fim de verificar a relevância temática do mesmo em relação a um assunto específico.

Este artigo está estruturado como segue: a seção 2 apresenta o referencial teórico; a seção 3, apresenta o MineraFórum e o seu algoritmo; a seção 4 descreve alguns detalhes da implementação da API. Na seção 5 é destacada a análise dos resultados obtidos e por fim, a seção 6 apresenta a conclusão e indicações de trabalhos futuros.

2. MineraFórum

O MineraFórum foi desenvolvido para verificar a relevância de uma postagem em relação a um assunto específico em fóruns de discussão. Esse recurso pode ser integrado a fóruns de discussão de ferramentas de ensino à distância, como por exemplo, o Moodle, para assim auxiliar os professores a avaliar a qualidade das postagens dos

alunos. O sistema foi desenvolvido utilizando a linguagem JAVA, buscando ser multiplataforma. O sistema baseou-se na técnica de mineração de textos implementada no software SOBEK e desenvolvido por Lorenzatti (2007), o qual permite gerar um grafo com informações obtidas do texto, formando uma rede de conceitos. Segundo Azevedo (2011) os grafos são construções matemáticas importantes para modelar o relacionamento entre conceitos. Outro motivo que influenciou na escolha por este software, é que o software SOBEK é um software livre.

Utilizando grafos para visualizar a saída de processamento produzida pela a mineração de textos é possível identificar quais palavras ocorrem com maior frequência nos textos e verificar se elas encontram-se próximas, já que para analisar o texto devem ser observadas as palavras e o seu contexto. Para o MineraFórum o contexto constitui-se no assunto tratado na discussão do fórum.

O MineraFórum utilizou em sua programação das seguintes funcionalidades do software Sobek: a) técnica de mineração utilizando grafos para visualização do resultado; b) definição do valor mínimo da frequência de um termo no texto; c) utilização de uma lista de stopwords; d) leitura de arquivos no formato txt, doc e pdf; e) possibilidade da criação de base de conceitos; f) leitura de uma base de conceitos gravados em um arquivo texto; g) construção do grafo; h) mineração dos textos em língua portuguesa ou inglesa; i) exibição dos recursos do sistema em língua portuguesa ou inglesa. Após utilizar o Sobek para gerar os grafos, Azevedo (2011), manualmente, fez um estudo para avaliar as postagens. O coeficiente de relevância temática (CRT) foi calculado segundo a equação (1):

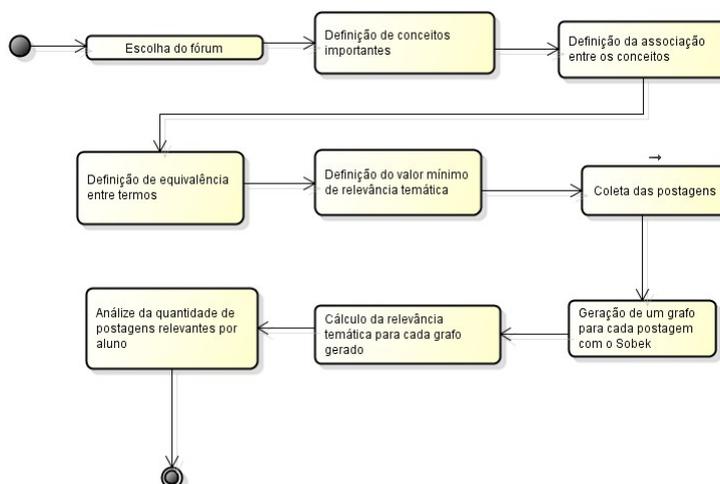
$$CRT = NC + NA \quad (1)$$

Onde,

- NC é o número de conceitos relevantes no texto;
- NA é o número de associações entre conceitos relevantes utilizadas no texto.

Azevedo (2011) seguiu a metodologia apresentada no diagrama de atividades da figura 1.

Figura 1 - Metodologia dos estudos preliminares



Para avaliar a solução proposta, Azevedo (2011) executou avaliações e testes. Analisando os resultados obtidos nos testes, foi possível comprovar que a mineração de textos utilizando grafos é uma opção adequada para a análise qualitativa das postagens em fóruns de discussão.

2.1 Algoritmo do MineraFórum

Na elaboração do algoritmo do MineraFórum, foram feitas entrevistas com 18 professores, sendo que essas entrevistas consistiram em dar uma nota de 0 a 10 para uma série de indicadores a serem utilizados na avaliação das postagens. Foram escolhidos os três indicadores que tiveram o maior peso e que poderiam ser implementados utilizando mineração de textos com grafos. Os três indicadores a seguir foram escolhidos para a avaliação: a) Verificar se a mensagem refere-se ao tema de debate, que teve média 9,61; b) A estruturação das postagens para verificar se uma postagem originou outras, que teve média 9,11; c) Verificação de postagens com textos muito parecidos, que teve média 7,61.

A Equação (2) foi utilizada por Azevedo (2011) para o cálculo de relevância das postagens,

$$RF = \frac{(P1.RT) + (P2.RM) - (P3.SM)}{P1 + P2 - P3} \quad (2)$$

Onde,

- RF representa a relevância da mensagem;
- P1, P2 e P3 representam o valor médio dos indicadores 1, 2 e 3, respectivamente.

Cabe ressaltar que, no cálculo de relevância temática, o MineraFórum consulta um dicionário de sinônimos e uma lista de equivalentes semânticos, mas é possível que existam termos importantes que não são contemplados. Sendo assim, com a união de todas as mensagens redigidas, o sistema realiza a mineração desse texto para tentar encontrar outros possíveis termos relevantes ao assunto. O grafo gerado a partir do fórum é utilizado no algoritmo. A equação (3) foi utilizada para o cálculo de relevância temática da mensagem.

$$RT = \frac{RT_1 + RT_2}{2} \quad (3)$$

Onde,

- RT1 representa a relevância temática da mensagem em relação ao texto de referência;
- RT2, que representa a relevância temática da mensagem em relação ao texto do fórum, foi utilizada a partir da descrição da equação (4):

$$RT_x = \left(RC + (DC * (1 - RC)) \right) + \left(PC * \left(1 - \left(RC + (DC * (1 - RC)) \right) \right) \right) \quad (4)$$

RC: Relevância dos conceitos da mensagem, equação (5).

$$RC = \frac{\sum_{i=1}^n f_i \cdot C_i}{\sum_{i=1}^N f_i \cdot C_i} \quad (5)$$

Onde,

$C_i =$ Nó equivalente;

f_i = Frequência do conceito C_i ;

$n =$ Total de nós equivalentes;

$N =$ Total de nós do grafo da mensagem.

Para RT1, o DC representa a distância relativa entre os conceitos relevantes da mensagem e o texto de referência e o PC é o peso relativo dos conceitos relevantes da mensagem e o texto de referência. Para RT2, o DC representa a distância relativa entre os conceitos relevantes da mensagem e o texto do fórum e o PC é o peso relativo dos conceitos relevantes da mensagem e o texto do fórum.

Para o cálculo da relevância das citações de uma mensagem foi utilizada a equação (6).

$$RM = \frac{m}{M} \quad (6)$$

Onde,

$RM:$ Relevância de citações da mensagem;

$m =$ quantidade de citações da mensagem m ;

$M =$ total de mensagens postadas no fórum.

O cálculo da similaridade entre as mensagens foi efetuado com a técnica de mineração de textos utilizando grafos, sendo que é feita uma comparação entre os grafos para determinar se as mensagens são semelhantes. Na equação (2), SM (Similaridade de uma mensagem com outra do fórum) recebe o valor zero se a mensagem não possuir similaridade com outra no fórum e recebe o valor de RT se a mesma possuir similaridade.

3. Protótipo e sua implementação

Nesta seção serão descritos os aspectos determinantes para o desenvolvimento da biblioteca de programação e também detalhes de sua implementação. O processo de desenvolvimento da biblioteca de programação começou com uma análise do algoritmo utilizado no MineraFórum. O algoritmo do MineraFórum utiliza três conceitos para realizar a sua análise, sendo que esses conceitos foram descritos na seção anterior. Os três aspectos básicos são: a) verificar se a mensagem refere-se ao tema de debate; b)

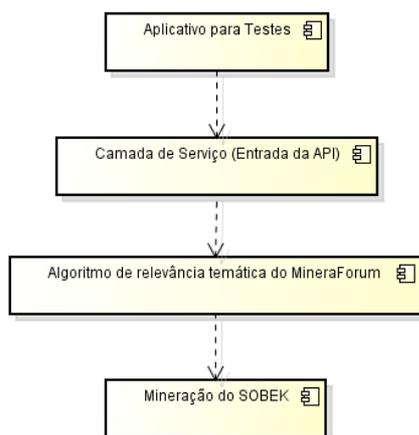
Analisar a estruturação das postagens para verificar se uma postagem originou outras; c)
Verificar a ocorrência de postagens com textos muito parecidos.

Logo ficou claro que, para o desenvolvimento da biblioteca de programação, seria necessário extrair apenas o primeiro conceito do algoritmo do MineraFórum, pois o objetivo principal é avaliar qualquer tipo de texto, principalmente artigos e construções mais extensas. O cálculo extraído do MineraFórum está descrito na equação 4 da seção anterior, que trata a relevância temática de um texto em relação ao texto base.

O funcionamento da biblioteca de programação está relacionado com as seguintes etapas: inicialmente é recebido um arquivo contendo um texto base, a partir do qual o algoritmo vai criar a base de conhecimento sobre o assunto a ser avaliado. Em seguida uma lista de textos a serem avaliados deve ser informada e então será iniciada a sua análise.

Na figura 2 é possível identificar uma visão geral da biblioteca de programação quanto a seus principais componentes.

Figura 2 - Visão geral da biblioteca de programação desenvolvida



Um aspecto importante da API é que o programador pode parametrizar o seu comportamento, definindo o conjunto de caracteres, a língua (que pode ser português ou inglês) e o mais importante, o valor mínimo que um termo deve aparecer no texto para ser considerado relevante. Esses termos formarão a base de conceitos que é utilizada na geração dos grafos.

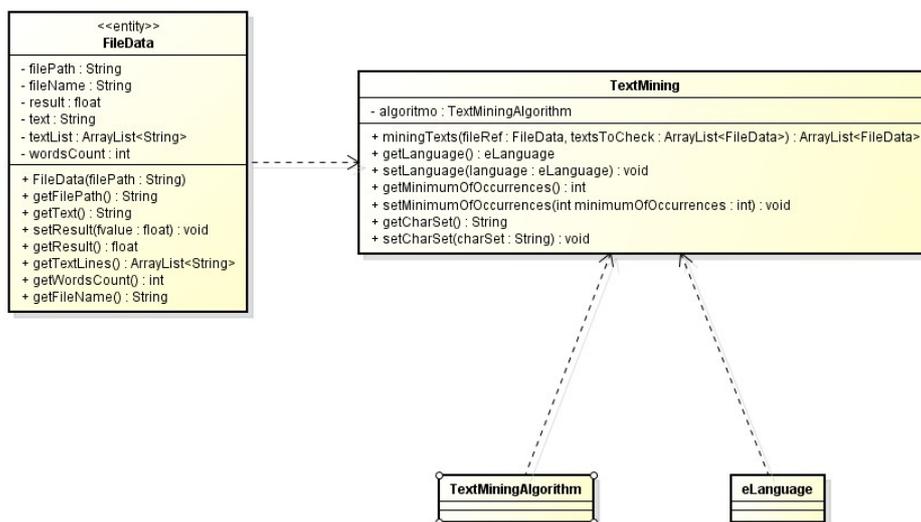
No processo de reengenharia de software, também verificou-se que o algoritmo não estava executando o processo de stemming para a língua inglesa, portanto foi adicionado o algoritmo de stemming de Porter¹, permitindo deste modo o uso da biblioteca tanto para textos em português como para textos em inglês.

Na figura 3 pode ser observado um diagrama de classes que representa os elementos principais do protótipo desenvolvido. As classes FileData e TextMining são

¹ <http://tartarus.org/martin/PorterStemmer/java.txt>

os componentes de acesso para a biblioteca de programação e representam a camada de serviço apresentada na figura 2. Já as classes eLanguage e TextMiningAlgorithm são as classes que, juntamente com o MineraFórum, compõe a camada do cálculo de relevância temática do MineraFórum.

Figura 3 - Classes da API desenvolvida



4 Método de Pesquisa

A proposta de trabalho adotada está embasada no estudo de um algoritmo para apoiar análise qualitativa em textos curtos e no desenvolvimento de uma API que possibilite a utilização dos aspectos principais deste algoritmo para um caso diferenciado, no qual serão analisados textos maiores.

A pesquisa constitui-se a partir de características aplicadas e exploratórias. A abordagem adotada é predominantemente qualitativa e tratou-se de um estudo de caso único. Durante o seu desenvolvimento privilegiou-se a construção de um protótipo a partir do qual foi possível a avaliação dos resultados e a verificação de relevância da proposta definida. A delimitação dos aspectos do estudo de caso empregado para a avaliação está ligada à características de conveniência no sentido de proporcionar o envolvimento dos profissionais necessários para suportar as medidas previstas.

5 Experimento e Resultados

Um aplicativo foi prototipado, visando permitir a avaliação da biblioteca de programação desenvolvida. A avaliação foi realizada com base em testes sobre as funcionalidades implementadas neste protótipo. Para os testes foi utilizado um computador com processador i5 com 4GB de memória no sistema operacional Windows 8.1. O aplicativo de testes, bem como a API, foram implementados na linguagem Java utilizando o SDK (*Software Development Kit*) 7 e a IDE (*Integrated Development Environment*) Netbeans 8.0.

O objetivo dos testes foi manter o cálculo realizado pelo MineraFórum para avaliar a viabilidade da sua utilização. Para o teste foi solicitado a ajuda de uma especialista

em biologia, graduada no ano de 2012 em licenciatura em ciências biológicas pela UCS - Campus Universitário da Região dos Vinhedos, para fornecer alguns artigos e o seu parecer sobre os mesmos. Quatro artigos foram fornecidos e testados. O primeiro foi considerado como o artigo base e os outros três artigos com certa similaridade de conteúdo. Junto a esses artigos foi adicionado ao teste um artigo que cujo conteúdo é totalmente diferente do conteúdo do texto base. Os parâmetros utilizados foram o uso de textos em língua portuguesa, a frequência mínima de termos igual a 6 e o tema dos artigos fornecidos para essa experimentação foi educação ambiental e estudo do solo.

A seguir são apresentadas algumas características do texto base usado como referência. O título do mesmo é “O solo como instrumento de educação ambiental” e a sua quantidade de palavras é igual a 4767. A execução da mineração retorna o valor do cálculo de acordo com a equação 5. Quando o texto é confrontado contra ele mesmo o resultado obtido foi igual ao valor 1,00 e o tempo de mineração foi de apenas 3 segundos.

A partir deste resultado foi possível avaliar os demais artigos. A seguir são descritos os resultados da mineração confrontando a base gerada do artigo de referência em relação aos demais quatro textos, sendo que a tabela 1 exhibe os resultados obtidos.

Tabela 1 – Resultados dos testes

Texto	Qtde. palavras	Tempo	Resultado
Texto 01	4925	00:00:03	1,00
Texto 02	4690	00:00:03	0,97
Texto 03	2065	00:00:02	0,00
Texto 04	6277	00:00:03	1,57

As bases de termos relevantes geradas pelo sistema para cada texto são exibidas na tabela 2, a seguir.

Tabela 2 – Termos relevantes gerados pelo protótipo

Texto	Termos relevantes
Texto base	agricultura, alunos, ambiente, bioturbação, campo, colorteca, cores, erosão, escola, forma, Fundamental, germinação, In, infiltração, LIMA, lugares, minhocário, pedológico, prática, região, Rondon-PR, SANTOS, SBCS, Solo, texturas, trigo, Viçosa,
Texto 01	agricultura, alunos, ambiental, assim, aula, Brasil, Cunha, degradação, escola, freático, Físico, geomorfologia, Guerra, hidrológico, humana, húmus, inadequada, intemperismo, Janeiro, LARACH, lixo, minerais, naturais, orgânica, outro, rios, rochas, social, sociedade, solo, superfície, tempo, Terra, vida, vivos, água.
Texto 02	Mestr, Ambient, ambiente, ANGELINI, anos, atividades, Biondi, COSTA, Curitiba, desenhos, Educ, escola, fauna, formal, fundamental, horizontes, In, LEPSCH, LIMA, mentais, MG, MUGGLER, Natureza, Paraná, Paulo, percepção, perfis, portanto, Rev, SOLO, tema, Universitária, água.
Texto 03	binomial, cara, coroa, cotidiano, Matemática, moedas, médio, Pascal, Probabilidades, viciadas.

Texto 04	ABREU, Alegre, ambiente, Anais, ar, aula, Bauru, BRASIL, Campinas, CARDOSO, CD-ROM, CEn, CI, ciclo, CIENCIAS, contexto, Curitiba, didáticos, eds, Educação, ensino, entanto, escola, exemplo, FALCONI, Fundamental, FURLANI, Geografia, geral, Goiânia, Horizonte, humano, IES, In, internet, licenciatura, Lima, Londrina, Maria, MEC, Miolo, Miolo , MONIZ, MOREAU, MUGGLER, médio, Nacionais, naturais, OLIVEIRA, Paraná, Paulo, PCN, pois, portanto, professor, publicações, Resumos, revista, SANTOS, SBCS, SEF, SIMP, solo, SUELO, tema, vida, vivos, Viçosa, água.
----------	---

Segundo a especialista, os textos 01 e 02 tem relevância temática muito próxima ao texto de referência, o que pode ser comprovado avaliando as bases de termos geradas pelo sistema e o resultado 1,00 e 0,97 respectivamente. O texto 03 foi inserido para verificar o resultado da mineração quando confrontado com um texto que trata de outro assunto, sendo que o resultado retornado pelo cálculo foi igual a 0. O resultado interessante nesse teste foi o valor retornado ao avaliar o texto 04. O texto 04 também foi fornecido pela especialista e na sua avaliação ela diz que esse texto fala sobre solo, mas não sobre educação ambiental, portanto a minha interpretação do resultado 1,57 é que o texto sim tem certa relevância, mas não total, pois o valor se afastou muito de 1,00 que seria o valor ideal. Outro ponto a se levar em consideração é que o texto 04 tem uma quantidade maior de termos que se repetem e que o algoritmo de mineração os trata como termos relevantes para o texto em questão.

6. Conclusão

Este artigo descreveu a implementação de uma biblioteca de programação para análise de textos, desenvolvido com base no algoritmo do sistema MineraFórum e a extração de parte da sua lógica. O objetivo desta integração foi empregar o algoritmo, que já havia comprovado bons resultados em textos curtos obtidos em fóruns de discussão e avaliar se este algoritmo pode ser utilizado para minerar textos maiores e de assuntos diversos.

O teste deste protótipo de API apresentou resultados interessantes, capazes de atestar que o algoritmo tem potencial para a sua utilização em construções textuais maiores do que postagens em fóruns de discussão. Outra questão a ser levada em conta é que as bases de termos geradas não são completamente adequadas, sendo que um trabalho grande deve ser feito em cima das StopWords e stemming para melhorar a mineração, além de entregar para o algoritmo uma base de termos mais consistente, como remover termos comuns em artigos e demais termos não relevantes.

Como trabalhos futuros, está prevista a atividade de integração desta biblioteca de programação com um buscador de textos na web, para verificar se o algoritmo também pode ser utilizado na avaliação de conteúdo da web, diretamente. Além desta atividade, novas etapas de testes com volumes maiores de textos estão previstas.

7 Referências

ALMEIDA, Siimone de; MARÇAL, Rui Francisco Martins; SCANDELARI, Luciano. Data Mining na Web para Inteligência Competitiva. XI SIMPEP – Bauru, SP, Brasil 2004. Disponível em <<http://www.pg.cefetpr.br/ppgep/Ebook/ARTIGOS/2.pdf>>.

- AZEVEDO, Breno Fabricio Terra. MineraFórum, Um recurso de apoio para análise qualitativa em fóruns de discussão. Porto Alegre, UFRGS, 2011.
- BARION, Eliana Cristina Nogueira; LAGO, Decio. Mineração de Textos. Revista de Ciências Exatas e Tecnologias, Vol. III N^a. 3, 2008. Disponível em: <<http://sare.unianhanguera.edu.br/index.php/rcext/article/view/413/409>>.
- LORENZATTI, A. SOBEK: uma Ferramenta de Mineração de Textos. 2007. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Departamento de Informática, Universidade de Caxias do Sul, Caxias do Sul, 2007.
- MORAIS, Edilson Andrade Martins; AMBRÓSIO, Ana Paula L. Mineração de Textos - Relatório Técnico. Universidade Federal de Goiás, 2007. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf>.
- MUKHERJEE, Indrajit; AL-FAYOUMI, Mohammad; MAHANTI, P.K.; JHA, Ritesh AI-BIDEWI, Ibrahim. Content Analysis based on Text Mining using Genetic Algorithm. 2nd International Conference on Computer Technology and Development (ICCTD 2010). Disponível em <<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5645835>>.
- Oliveira, R. L. J., Esmín, A. A. A. 2012, Monitoramento Automático de Mensagens de Fóruns de Discussão Usando Técnica de Classificação de Texto Semi-Supervisionado. Anais do 23^o Simpósio Brasileiro de Informática na Educação (SBIE 2012), ISSN 2316-6533 Rio de Janeiro, 26-30 de Novembro de 2012.
- RIGO, S. J. ; Cambuzzi, w. ; Barbosa, J. L. V. ; CAZELLA, Sílvio . Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. Revista Brasileira de Informática na Educação, v. 22, p. 132-146, 2014.
- Silva, J. K.K., Bastos, H. P. P., Bercht, M., Wives, L. K. 2012. Automatização do Processo de Identificação de Presença Social em Fóruns e Chats . Anais do 23^o(SBIE 2012, ISSN 2316-6533 Rio de Janeiro, 26-30 de Novembro de 2012
- SCHELLER, Thomas; KUHN, Eva. Influencing Factors on the Usability of API Classes and Methods. 2012 19th IEEE International Conference and Workshops on Engineering of Computer-Based Systems. Disponível em <<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6195191> >.
- TAN, A. Text Mining: The State of the Art and the Challenges. Kent Ridge Digital Labs 21Heng Mui Keng Terrace, Singapore, 1999. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.132.6973>>.