Análise de Ferramentas de Mineração de Textos para Apoio à Produção Textual

Miriam Klemann[1] , Eliseo Reategui[1,2], Clevi Rapkiewicz[1]

[1] PPGEDU, [2] PPGIE - UFRGS, Av. Paulo Gama, 110 - Porto Alegre/RS - Brazil

{miriamklemann, eliseoreategui, clevirap}@gmail.com

Resumo: *Este artigo apresenta um estudo comparativo sobre diferentes ferramentas de análise e mineração de textos, tendo-se como princípio a utilização destas como recurso educacional. Quatro ferramentas são analisadas a partir de diferentes critérios, tais como: facilidade de operação, visualização de termos relevantes e disponibilidade na web.*

Abstract: *This paper presents a comparative study about different tools for text analysis and mining, having as a main principle their use as na educational resource. Four tools have been analysed using different criteria: ease of use, visualization of relevant terms and online availability.*

1. Introdução: mineração de textos

A mineração de textos pode ser definida como um método de extração de informações relevantes em bases de dados não estruturadas, ou semi-estruturadas (FELDMAN e SANGER, 2006). Trata-se de um campo multidisciplinar que inclui conhecimentos de áreas como Informática, Estatística, Linguística e Ciência Cognitiva. A mineração de textos busca extrair regularidades, padrões ou tendências de textos em linguagem natural, normalmente, para objetivos específicos. Dentre outros, a área tem despertado muito interesse em decorrência da popularidade da Internet, da geração e fácil acesso a vastos repositórios de textos (SHARP, 2001). Outra definição é a de Mattison (1999) para quem mineração de textos é uma aplicação de sistemas de computação que envolve *hardware* e *software* dedicados à análise textual de documentos. A técnica pode ser vista como uma extensão da área de *Data Mining*, cujo foco é na análise de dados estruturados. Também chamada de *Mineração de dados textuais* ou *Descoberta de Conhecimento em Textos*, a mineração de textos permite recuperar informações, extrair dados, resumir documentos, descobrir padrões, dentre outras análises possíveis de se realizar em documentos de texto. Pode ser utilizada com muitos propósitos, como por exemplo identificar documentos similares entre si, buscar dados relevantes dentro do documento, entre outras.

O foco deste artigo está na mineração de texto como apoio a produção textual. Isso porque ferramentas de mineração de texto podem fornecer pontos positivos e/ou negativos sobre um texto permitindo a identificação de alguns problemas, tais como: a necessidade de explorar mais um determinado tema, a necessidade de produzir um texto mais fluido, e não apenas uma justaposição de termos que não estão bem conectados (KLEMANN et al, 2009). Uma questão importante é a escolha da ferramenta mineradora a utilizar. Que recursos elas devem apresentar para serem utilizadas para apoio a produção textual? Com base nesse questionamento, analisamos algumas

ferramentas e comentamos as possibilidades de uso das mesmas. Para tanto, organizamos este artigo em 3 seções além da presente introdução. Na seção 2, discutimos os recursos que as ferramentas devem apresentar. Na seção 3, apresentamos as ferramentas analisadas e os respectivos recursos. Finalmente, apresentamos algumas considerações finais.

2. Recursos importantes para apoio a produção textual

A tarefa de escrita nem sempre é uma tarefa simples para aqueles que escrevem, podendo implicar dificuldades de diferentes naturezas. Ao mesmo tempo, o avanço da tecnologia, exige que o professor esteja continuamente pesquisando para que possa desenvolver e criar atividades desafiadoras para seus alunos utilizando recursos e metodologias adequadas. Alguns recursos foram analisados: [i] disponibilidade ***online***: tem um papel importante. Oferecer ambientes para que o aluno acesse a qualquer momento e possa desenvolver seu conhecimento. [ii] **contagem de termos** de um texto: facilita nas produções de texto que possuem limite de palavras estabelecidas. [iii] **apresentação de todos os termos**: permite visualizar os termos bem como possíveis repetições de alguns deles. [iv] **apresentação de termos relevantes**: permite ao usuário descobrir rapidamente o assunto principal do texto analisando as palavras e expressões utilizadas. [v] **freqüência dos termos**: refere-se a ocorrência de um termo no texto (variável, pois o próprio usuário pode estabelecer). [vi] **relacionamentos dos termos:** fundamental para estruturação das idéias, importante do planejamento e preparação para a tarefa de escrita do texto. [vii] **visualização gráfica dos termos**: facilita a observação dos termos centrais de um texto. [viii] recurso **visualização gráfica dos termos e relacionamentos**: a partir da observação destes termos, o usuário é capaz de delinear similaridades e diferenças entre os conceitos, podendo desenvolver compreensões mais precisas e integradas e assim reformular novas proposições usando suas próprias palavras (AUSUBEL, 1982).

3. Ferramentas analisadas

A seguir, algumas ferramentas para a análise de texto são apresentadas.

TextAlyser[1] é uma ferramenta de análise de textos *online* gratuito utilizada para destacar grupos de palavras-chave. A utilização desta ferramenta permite ao usuário descobrir rapidamente o assunto principal do texto analisando as palavras e expressões utilizadas. Inicialmente a ferramenta faz uma contagem dos termos utilizados no texto apontando o número total de palavras e apresentando uma série de estatísticas sobre palavras e termos mais frequentes. Analisa a complexidade e capacidade de leitura de qualquer texto ou *website*. O programa aponta também a frequência com que as palavras mais utilizadas ocorrem no texto, bem como número de palavras, número de sílabas, dentre outros. Além destes fatores, a ferramenta ainda apresenta um índice relativo à "facilidade de leitura" (*readability*), critério obtido a partir do tamanho das frases e estatísticas encontradas. O programa não apresenta nenhuma ferramenta gráfica para visualização das principais informações contidas no texto.

[1] Software disponível em: <http://textalyser.net>. Acesso em: julho 2011.

Wordcounter[2] foi criado por Steven Morgan Friedman. É uma ferramenta *online* gratuita que apresenta a relação das palavras mais utilizadas em um texto. Para os autores de textos é útil, pois mostra as palavras repetidas e/ou redundantes numa lista. Nesta lista constam quais palavras foram utilizadas e a frequência com que cada uma aparece no texto. O programa Wordcounter também tem como principal objetivo encontrar estatísticas relativas ao uso de palavras e termos no texto, não dispondo de ferramentas gráficas mais complexas para visualização das informações.

TagCrowd[3] é uma ferramenta online criada por Daniel Steinbock (*Stanford University California - North America*), que permite criar nuvens de marcadores (*tagclouds*) de qualquer texto, em diferentes línguas. Para operar a ferramenta, inicialmente estabelece-se qual a frequência mínima que a palavra deve aparecer no texto e também o número de palavras que o programa terá que mostrar. Após a definição destes parâmetros, pode-se visualizar os termos do texto em vários formatos e cores diferentes. Diferentemente das ferramentas descritas anteriormente, esta apresenta os principais termos do texto de uma forma mais visual, enfatizando palavras e termos mais frequentes. A ferramenta não busca encontrar relações entre os termos, sendo estes apresentados em ordem alfabética.

A ferramenta **Sobek**, desenvolvida por uma equipe multidisciplinar na Universidade Federal do Rio Grande do Sul, pode ser executada em computadores com diferentes sistemas operacionais Linux, Windows ou Mac OS, permitindo que seja utilizada sem maiores restrições. Contudo não está disponível *online*. A ferramenta é capaz de minerar textos em diversos formatos (txt, doc, pdf) de maneira bastante rápida, possibilitando seu uso efetivo em situações de ensino reais. Com relação a outras ferramentas de análise de textos apresentadas, a ferramenta Sobek se distingue principalmente por apresentar tanto os principais conceitos encontrados em um texto, quanto os relacionamentos entre estes, empregando grafos (REATEGUI et al, 2011). A tabela 1 compara diferentes recursos dos softwares analisados.

Tabela 1 – Análise comparativa de Mineradores

	Online	Contagem de termos	Apresenta-ção de todos os termos	Apresenta-ção de termos relevantes	Freqüência dos termos	Relacio-namentos dos termos	Visualiza-ção gráfica dos termos	Visualização gráfica dos termos e relacionamentos
TextAlyser	X	X			X			
Wordcounter	X	X	X		X			
TagCrowd	X			X			X	
Sobek				X	X	X	X	X

As ferramentas *TextAlyser* e *Wordcounter* apresentam uma série de estatísticas relativas ao uso de palavras e termos mais frequentes de um texto. Já a ferramenta *TagCrowd*, realiza somente uma extração de frequência de termos em um texto, sendo estes apresentados em ordem alfabética numa nuvem de marcadores. A ferramenta

[2] Disponível em: <http://www.wordcounter.com/>. Acesso em: julho 2011.

[3] Disponível em: <http://tagcrowd.com/>. Acesso em: julho 2011.

Sobek possui características distintas das demais ferramentas apresentadas, pois tem uma maior capacidade de extrair informações relevantes. Relaciona conceitos que são obtidos por uma análise da distância entre os termos de um texto. Quanto mais próximos dois termos estiverem em um texto, mais relacionados estarão.

4. Considerações finais

A análise dos vários recursos permite destacar especial importância para as ferramentas *online*. Considerando a atual tendência de computação em nuvem, ferramentas desktop são menos flexíveis. Por outro lado, esta flexibilidade pouco acrescenta para o usuário se outras caracaterísticas importantes estiverem ausentes, particularmente a possibilidade de visualização de relacionamento entre os termos e respectiva visualização gráfica. Ainda, a possibilidade de manipular o grafo apresentado diretamente na forma gráfica é importante. As estatísticas apresentadas pelas ferramentas *TextAlyser, Wordcounter* e *TagCrowd* podem contribuir na produção, revisão e avaliação de textos. A ferramenta *Sobek* apresenta resultados mais completos, além de ser de fácil manejo. A partir dos termos do grafo, os alunos podem visualizar e esquematizar mentalmente a seqüência dos conceitos e suas relações, e assim, formular frases organizando a sua produção textual. A ferramenta de mineração de textos *Sobek* deverá ser disponibilizada para outras aplicações e outras instituições, buscando compreender como a extração automática de conceitos e sua representação gráfica pode contribuir em outras atividades educacionais, como por exemplo, em processos de letramento e resolução de problemas.

Agradecimentos

Esta pesquisa foi parcialmente financiada pelo CNPq, processo 476398/2010-0, e FAPERGS, processo 1018248.

Referencias:

Ausubel, D. P. (1982) **A aprendizagem significativa: a teoria de David Ausubel**. São Paulo: Moraes..

Feldman, R., Sanger, J. (2006). **Text Mining Handbook**. Cambridge, Inglaterra: Cambridge University Press.

Klemann, M. et al. (2009). O Emprego da Ferramenta de Mineração de Textos SOBEK como Apoio à Produção Textual. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 20º , Florianópolis. **Anais**. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/1154/1057>. Acesso em: junho. 2010.

Mattison, R.; Mattison, B. K. (1999). **Web Warehouse and Knowledge Management**. New York: Mc Graw Hill..

Reategui, E. et al. (2011). **Sobek: a text mining tool for educational applications**. In: INTERNATIONAL CONFERENCE ON DATA MINING, Las Vegas, Nevada, USA. **Anais**. Las Vegas: [s.n.], 2011. p. 59-64.

Sharp, M. (2001). **Text Mining, Rutgers University, School of Communication, Information and Library Studies**. Disponível em: <http://www.scils.rutgers.edu/msharp/textminig. htm>. Acesso em: junho. 2010.