Uso de Grafos de Termos para Análise do Conteúdo de Documentos Técnicos

Luiz Cláudio S. Silva, Renelson R. Sampaio

Faculdade de Tecnologia SENAI Cimatec
Av. Orlando Gomes, 1845 - Piatã – 41.650-010 – Salvador – BA – Brazil

{luizclaudio,renelson.sampa}@gmail.com

Abstract. *Visualization methods for isolated documents are either too simple, based in frequency of terms only, or depend on syntactic and semantic information bases to be more significant. This paper proposes an intermediary method, based on automatic summarization algorithms, and tries to bring more information without using external sources. The method considers the frequency of repetition of pairs of important terms and the distance between them.*

Resumo. *Métodos de visualização de documentos isolados geralmente são bem simples, baseados somente na frequência de termos, ou dependem de bases de informações sintáticas e semânticas para serem mais significativos. Este trabalho propõe um método intermediário, baseado em algoritmos de resumo automático de texto, e tenta agregar mais informação sem necessitar de dados externos. São consideradas a frequência de repetição de pares de termos importantes e a distância entre eles a cada encontro.*

1. Introdução

Diferentemente de quando aplicada a conjuntos de documentos, a visualização de conteúdo apresenta dificuldades particulares ao ser utilizada em documentos isolados. No primeiro caso, métodos estatísticos podem ser mais largamente empregados graças à maior quantidade de informação disponível. Já no segundo, que muitas das vezes trata de textos curtos, a falta de dados torna necessária a análise das estruturas sintáticas e semânticas do documento [Grobelnik 2004].

Analisar estruturas sintáticas e semânticas de documentos isolados requer utilizar bases de palavras com categorias gramaticais anotadas ou organizadas semanticamente. Um exemplo deste tipo de visualização são as redes semânticas. Caso essas bases de palavras não estejam disponíveis, ainda é possível gerar visualizações baseadas somente na frequência de ocorrência das palavras como, por exemplo, uma nuvem de termos (*tag cloud*).

2. Objetivo

O presente trabalho tem como objetivo propor uma forma de visualização de documentos isolados que possa apresentar mais informação que uma simples nuvem de termos mas, ao mesmo tempo, não necessite se utilizar de bases de palavras

previamente processadas. A motivação para tanto é a possibilidade de, com um pouco mais de esforço não supervisionado, conseguir um resultado que permita ter uma visão mais esclarecedora sobre o texto.

3. Metodologia

A forma de visualização aqui proposta se baseia no princípio adotado por algoritmos de geração automática de resumos, particularmente o de [Luhn 1958], no qual o ou os assuntos principais de que trata um documento podem ser determinados através da análise dos termos que mais ocorrem no mesmo. As sentenças mais importantes, que são usadas para compor o resumo do texto, são aquelas em que os termos mais frequentes aparecem em maior quantidade e mais próximos. Como muitos termos aparecem mais de uma vez, é adotado um critério de corte de só considerar aqueles que se repitam mais que a média de repetições no documento. Caso o número de termos selecionados ainda seja muito grande para fins práticos, o valor de corte pode ser aumentado para média de repetições mais meia ou uma vez o desvio padrão, por exemplo.

O que se propõe é identificar os termos mais importantes do texto e as possíveis relações entre eles. Seguindo procedimentos padrão de mineração de texto [Weiss 2005], o documento deve ser pré-processado para a eliminação de palavras negativas (*stopwords*), redução dos termos a seus radicais (*stemming*) e identificação de expressões (*n-grams*). Então, são identificados os termos mais frequentes e levantados o número de vezes em que ocorrem nas mesmas sentenças e a distância entre eles a cada encontro. A médias de cada um destes dois fatores, frequência de ocorrências conjuntas de pares de termos e distância entre os termos do par nas sentenças, são normalizadas em relação a todos os pares identificados no documento, de modo que os maiores valores passam a ser considerados 1,0 e os demais são escalonados proporcionalmente até zero. Os valores normalizados de média da frequência e média da distância de cada par são multiplicados entre si, gerando um terceiro valor que indica a força do vínculo entre os dois temos, no contexto do documento. Este valor, também varia de zero a 1,0.

A partir dos dados encontrados é criado um grafo onde os vértices são os termos mais importantes do texto e as arestas são traçadas caso haja vínculos entre eles. A análise visual do grafo permite avaliar conceitos como centralidade, intermediação e proximidade entre os termos, o que ajuda a entender a estrutura do documento. Partes desconectadas podem significar, por exemplo, assuntos diferentes tratados no mesmo documento ou problemas no estilo do redator.

Para agregar mais à visualização do documento, propõe-se ainda desenhar os vértices em tamanho proporcional ao número de ocorrências do termo no documento e utilizar o valor calculado do vínculo entre os termos como espessura das arestas que os unem. Ou seja, vértices maiores denotam palavras mais frequentes e arestas mais grossas vínculos mais fortes.

4. Resultados

Para validar a metodologia proposta foram utilizados documentos produzidos pelo Serviço Brasileiro de Respostas Técnicas (SBRT), uma iniciativa do Instituto Brasileiro

de Informação em Ciência e Tecnologia (IBICT), órgão do Ministério da Ciência, Tecnologia e Inovação (MCTI), que, desde 2005, atende demandas de empresários e empreendedores por informação técnica e tecnológica de baixa complexidade sobre produtos e processos, com o objetivo de auxiliá-los a se estabelecer ou melhorar seus negócios [SBRT]. A decisão de usar documentos do SBRT se deveu à qualidade da redação dos mesmos e à diversidade de assuntos de que tratam. Foi utilizado um subconjunto de 55 dossiês técnicos produzidos até abril de 2012 pelo Instituto Euvaldo Lodi – Regional Bahia (IEL/BA), instituição participante do SBRT. Cada documento foi obtido como um arquivo do tipo Portable Document Format (PDF) e então convertido para o formato texto com codificação Unicode de 8 bits (UTF-8), que dá suporte aos caracteres acentuados da Língua Portuguesa. Na média, os documentos têm 5.800 palavras. Para as tarefas de mineração de texto e visualização foram utilizadas rotinas escritas na linguagem de programação Python [Python] incorporando recursos das bibliotecas *Natural Language Toolkit* [NLTK], para processamento de linguagem natural e *Simple Graphics Library* [Zelle 2010], para desenho. Na figura 3 pode se visto um grafo gerado a partir do DT 257 do SBRT, cujo título é "Avaliação dos agentes químicos na Construção Civil".

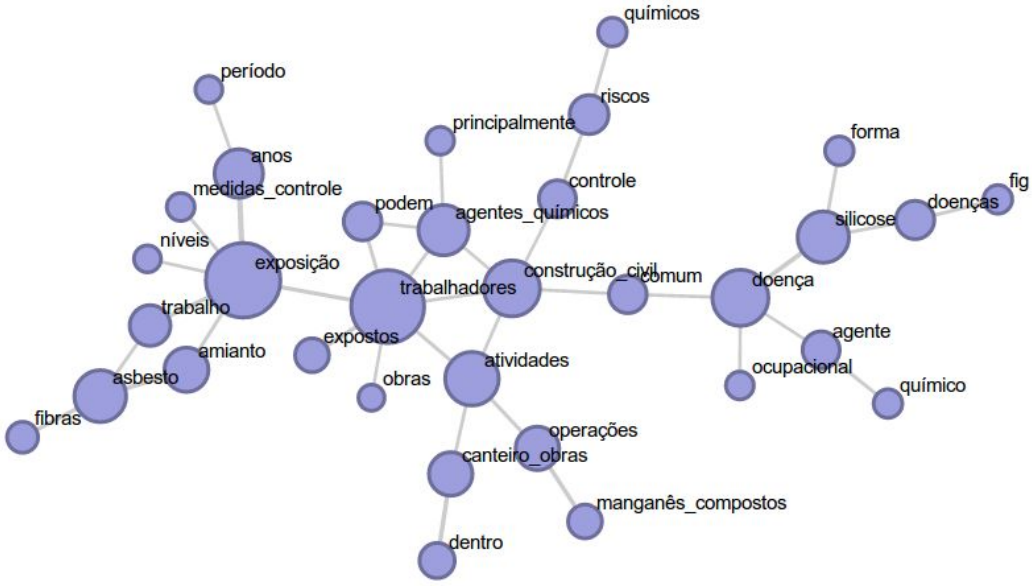

Figura 3. Grafo gerado a partir do DT 257.

Referências

Grobelnik, M., Mladenic, D. (2004). Tutorial on Text Mining em *PASCAL Network of Excellence Workshop on Text Classification*

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159–165.

NLTK - Natural Language Toolkit, Disponível em: http://www.nltk.org/, Acesso em: 14 abril 2012.

Python Programming Language – Official Website, Disponível em: http://python.org/, Acesso em: 14 abril 2012.

SBRT, Disponível em: http://www.respostatecnica.org.br/, Acesso em: 14 abril 2012.

Weiss, S. M., N. Indurkhya, T. Zhang, & F. J. Damerau. (2005). From textual information to numerical vectors. In Text Mining: Predictive Methods for Analysing Unstructured Information, pp. 15–44. Springer Verlag

Zelle, J. M. (2010). Simple Graphics Library, Disponível em: http://mcsp.wartburg.edu/zelle/python/graphics/graphics/index.html