MINERAÇÃO DE TEXTOS: CONCEITO, PROCESSO E APLICAÇÕES

Anderson Pezzini

ander_pezzini@hotmail.com

Resumo
A mineração de textos é uma extensão da mineração de dados, e pode ser definida como um processo de extração de informações desconhecidas e úteis de documentos textuais escritos em linguagem natural. Como a maioria das informações são armazenadas em forma de texto, a mineração de textos possui alto valor comercial, e pode ser aplicada em áreas como medicina e atendimento ao cliente.

Palavras-Chave: Mineração de Dados; Mineração de Textos.

Abstract
The text mining it's a extension of data mining, and can be definedas a process of extraction of unknown and useful information from textual documents written in natural language. As most of the information is stored in text form, the text mining has a high commercial value, and can be used in areas like medicine and customer service.

Keywords*: Data Mining; Text Mining.*

1. Introdução

Graças ao desenvolvimento da Internet e das redes de computadores os documentos virtuais se transformaram no principal método de armazenamento de informações, principalmente as informações comerciais, que segundo estimativas armazena cerca de 85% de suas informações em documentos de texto. Porém, os paradigmas mais tradicionais de desenvolvimento de software não são capazes de entender o relacionamento confuso e geralmente ambíguo que existe nos documentos de texto virtuais (MACHADO et al., 2010).

A mineração de textos é um paradigma de programação criado para resolver este problema, sendo capaz de entender a linguagem natural dos documentos de texto e conseguindo lidar com a sua imprecisão e incerteza. A mineração de textos envolve várias áreas da informática, como mineração de dados, aprendizado de máquina, recuperação de informação, estatística e linguagem computacional, para conseguir transformar o texto em algo que um computador consiga entender (MACHADO et al., 2010).

O principal objetivo da mineração de textos é encontrar termos relevantes em documentos de texto com grande volume de dados e estabelecer padrões e relacionamentos entre eles com base na frequência e temática dos termos encontrados (SERAPIÃO, 2010).

A tecnologia de mineração de textos não é um mecanismo de busca, pois a mineração ajuda o usuário a descobrir informações previamente desconhecidas, enquanto na busca o usuário já sabe o que deseja procurar. Além disso, a mineração

também é diferente de robôs de conversação (*chatterbot*), pois ela não tenta simular o comportamento humano (ARANHA; PASSOS, 2006).

2. Técnicas utilizadas

As principais técnicas utilizadas para fazer a mineração de textos são:

Processamento de Linguagem Natural: É um método que procura utilizar computadores para melhorar o entendimento da linguagem natural através de técnicas para processar textos rapidamente, utilizando-se de manipulação de strings até linguagem natural de inquéritos (MACHADO et al., 2010).

Recuperação de Informação: Utiliza métodos e medidas estatísticos ou semânticos para automaticamente processar o texto de documentos para encontrar quais documentos possuem a resposta para a questão (mas não a resposta em si). Embora já fossem utilizadas técnicas deste tipo de forma primitiva em 1975, este método só ganhou notoriedade com a popularização da Internet(MACHADO et al., 2010).

Extração de Informação: Possui como principal objetivo buscar partes relevantes de um texto em um documento e extrair informações específicas destas partes. Possui um conceito mais limitado da compreensão da linguagem natural(MACHADO et al., 2010).

Estas técnicas são vastamente utilizadas na mineração de dados, principalmente em redes sociais e em processos de ensino a distância (MACHADO et al., 2010).

3. O processo de mineração de dados

A mineração de textos pode conter várias etapas, mas quatro delas são básicas em todos os processos: coleta de documentos, pré-processamento, extração de conhecimento e avaliação e interpretação dos resultados (MARTINS et al., 2003).

A coleta de documentos objetiva conseguir documentos relacionados ao tipo de conhecimento que se deseja obter. Pode-se utilizar de várias fontes, como livros, e-mails, fóruns de internet, etc. Existem técnicas como o Processamento de Linguagem Natural e Recuperação de Informação que podem ser utilizadas nesta etapa (MARTINS et al., 2003).

Uma característica importante deste passo é a limitação quanto ao uso de informações externas. O conhecimento de mundo e de especialistas não são utilizados pois os algoritmos de agrupamento dos documentos aprende de forma não supervisionada como extrair o conhecimento dos textos, e desta maneira, categorizar os documentos de forma a facilitar os próximos passos do processo de mineração de textos (CORRÊA et al., 2012).

No pré-processamento os documentos adquiridos na primeira etapa, escritos em linguagem natural, passam por uma formatação para estrutura-los de maneira padronizada, mas sem perder suas características naturais, para que os algoritmos que serão utilizados nas próximas etapas sejam capazes de manipular todos os documentos da mesma maneira. Ao final deste processo obtém-se uma estrutura que representa o grupo de documentos fonte, geralmente uma tabela atributo-valor (MARTINS et al., 2003).

Neste segundo passo também são definidos os termos que serão utilizados para a extração para conseguir um grupo pequeno e representativo dos termos presentes nos

grupos de textos. Para isto, eliminam-se as *stopwords*, termos sem significado relevante para a pesquisa, como artigos, advérbios e pronomes. Além disso, são identificadas variações morfológicas e sinônimos dos termos, utilizando técnicas como *stemming* e *thesaurus*, a fim de diminuir ainda mais o conjunto de pesquisa. Esta redução permite a diminuição do custo computacional das próximas etapas do processo (CORRÊA et al., 2012).

Na terceira etapa, a extração do conhecimento, aplica-se algoritmos de extração automática de conhecimento para buscar informações desconhecidas até o momento, mas que possam ser úteis para o domínio da questão (MARTINS et al., 2003).

Estes algoritmos buscam agrupar objetos similares através de uma medida de proximidade, ao mesmo tempo em que tenta formar grupos com características dissimilares entre si. A análise por agrupamento também é chamada de análise exploratória de dados ou aprendizado por observação (CORRÊA et al., 2012).

Por último, na avaliação e interpretação dos resultados, utiliza-se a ajuda de um usuário para descobrir se os resultados obtidos são satisfatórios, e em caso negativo, que etapas poderiam ser refeitas para melhorá-los (MARTINS et al., 2003).

4. Aplicações

A mineração de textos possui aplicações nas mais variadas áreas científicas e comerciais. Ela pode ser usada, por exemplo, na medicina, pois a quantidade de informação de texto gerada nesta área é enorme (prontuários, registros hospitalares, fichas de pacientes, etc). Estes documentos podem ser avaliados com técnicas de mineração de dados para auxiliar os profissionais da medicina a diagnosticar doenças ou buscar tratamentos. Uma ferramenta que utiliza este conceito é o Medline, um software que trabalha coma base de dados bibliográficos da Biblioteca Nacional de Medicina dos Estados Unidos da América (CARRILHO JUNIOR, 2007).

Além disso, a mineração de textos pode ser utilizada para a análise de sentimentos em pesquisas de opinião pública. Muitas vezes estas pesquisas são feitas com questionários com perguntas fechadas, ou seja, os entrevistados podem escolher somente opções pré-determinadas. O problema é que muitas vezes isto não reflete a realidade, pois as perguntas podem exigir uma respostas mais elaborada. Se em fez deste tipo de questionário for utilizada uma entrevista com respostas abertas, de maneira que o entrevistado possa escrever sua resposta em linguagem natural, é possível analisar os resultados com uma ferramenta de mineração de textos (CARRILHO JUNIOR, 2007).

Esta área da mineração de textos é conhecida como “Análise de Sentimentos”, e visa identificar como o autor de um texto expressa seus sentimentos de forma escrita e categorizar a satisfação (favorável ou desfavorável) em relação ao assunto abordado(CARRILHO JUNIOR, 2007).

Por último, a mineração de textos pode ser utilizada para ajudar empresas grandes que trabalham com atendimento ao cliente. Muitas vezes, um produto ou serviço apresenta algum defeito e o cliente precisa entrar em contato com algum especialista da empresa para resolver o seu problema. É comum nestes casos a requisição do cliente ser transferida de setor para setor e demorar muito tempo até chegar ao seu destino final (CARRILHO JUNIOR, 2007).

A proposta da mineração de textos para solucionar este problema é analisar textualmente a requisição do cliente e enviá-la de maneira automática diretamente para

o especialista no assunto, removendo a intervenção humana do processo e aumenta o tempo de entrega da requisição (CARRILHO JUNIOR, 2007).

5. Considerações Finais

A mineração de textos possui potencial para ser muito bem explorada comercialmente, não apenas pela grande variedade de informações comerciais que são armazenadas em forma de texto, mas também por causa da sua capacidade de ser aplicada em várias áreas diferentes do conhecimento.

Além disso, a tecnologia da mineração de dados pode fornecer automação para vários serviços que atualmente são feitos por seres humanos e, consequentemente, a diminuição de custo e maior agilidade nestes processos. Exemplos disto são a análise automática de sentimentos em pesquisas de opinião pública, que quando realizadas por pessoas demoram muito mais tempo do que se realizadas por um algoritmo de mineração de dados.

Por fim, é importante salientar que mesmo com toda a automação fornecida pela mineração de dados, ainda é necessário alguém ao final do processo para avaliar os resultados obtidos na mineração.

Referências

ARANHA, Christian; PASSOS, Emmanuel. **A Tecnologia de Mineração de Textos.** 2006. Disponível em: <http://189.16.45.2/ojs/index.php/reinfo/article/view/171>. Acesso em: 01jun. 2015.

CORRÊA, Geraldo Nunes et al. **Uso da mineração de textos na análise exploratória de artigos científicos.** 2012. Disponível em: <http://www.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_383.pdf>. Acesso em: 01 jun. 2015.

CARRILHO JUNIOR, João Ribeiro. **Desenvolvimento de uma Metodologia para Mineração de Textos**. 2007. Disponível em: <http://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=11675@1>. Acesso em 02 jun. 2015.

MACHADO, Aydano P. et al. **Mineração de Texto em Redes Sociais Aplicada à Educação a Distância.** 2010. Disponível em: <http://pead.ucpel.tche.br/revistas/index.php/colabora/article/view/132>. Acesso em: 01 jun. 2015.

MARTINS, Claudia Aparecida et al. **Uma Experiência em Mineração de Textos Utilizando Clustering Probabilístico Clustering Hierárquico.** 2003. Disponível em: <http://www.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_205.pdf>. Acesso em: 01 jun. 2015.

SERAPIÃO, Paulo Roberto Barbosa et al. **Uso de mineração de texto como ferramenta de avaliação da qualidade informacional em laudos eletrônicos de mamografia.** 2010. Disponível em: <http://www.scielo.br/pdf/rb/v43n2/a10v43n2.pdf>. Acesso em: 01 jun. 2015.