Mineração de Texto

Marcos Cordeiro de Brito Jr
Programa de Pós Graduação em Engenharia Elétrica e Computação (PPGEEC)
Universidade Presbiteriana Mackenzie
São Paulo, Brasil

***Resumo*—Este documento é um resumo sobre a técnica de Mineração de Texto. Será apresentado sua definição, processos, técnicas e aplicações práticas dos estudos.**

***Index Terms*—Mineração de Texto, Mineração de Dados, Conhecimento**

I. Introdução

A técnica de mineração de texto tem por objetivo extrair informações analisando grandes volumes de dados não estruturados e semi-estruturados utilizando algorítimos computacionais. Essas análises buscam padrões, agrupamentos, regularidades e tendências para conseguir identificar conhecimentos que possam trazer algum valor para a análise.

Qualquer área pode se beneficiar das técnicas de mineração de texto para conseguir compreender melhor o valor de grandes volumes de informações, onde as formas comuns de buscas ou leitura do conteúdo se tornam inviáveis nas maneiras convencionais.

Esse documento tem por objetivo descrever de forma sucinta e resumida a técnica de mineração de texto e sua organização está disponível da seguinte maneira: A Seção II descreve de maneira geral, através de referências acadêmicas, as definições sobre mineração de texto. Na Seção III são mostrados as etapas do processo e algumas técnicas e na Seção IV, algumas aplicações práticas dessa técnica.

II. Definição de Mineração de Texto

Mineração de textos é um conjunto de métodos usados para navegar, organizar, achar e descobrir informação em bases textuais [1].

O termo vem ganhando popularidade durante os anos conforme o número de pesquisas avançam. Em um estudo publicado em 2015 pela *Emerald Insight* foi analisado a quantidade de citações e estudos sobre o termo *Text Mining (Mineração de Texto)* na base de dados *Web of Science* do Departamento e Biblioteca de Ciência da Informação da Universidade de *Savitribai Phule Pune* na *India* [2].

Durante os anos de 1999 e 2013 houveram 8102 citações do termo em 987 publicações. Entre os países com maiores contribuições estão os Estados Unidos com 384 *papers* publicados, seguido da China com 98 e Taiwan com 74. Comparando entre as universidades com os maiores números de pesquisas estão: *Columbia University* e *Univerity Arizona* dos Estados Unidos e *Nanyang Technological University* de Singapura.

Com o grande aumento do volume de dados e com constante crescimento a cada ano, impulsionado principalmente pelas redes sociais, a descoberta de conhecimento tem se tornado cada vez mais difícil e cada vez mais importante. Para auxiliar nessa exploração, surgiram técnicas de análise de dados que ajudam a navegar nesse oceano infinito de informações.

Uma dessas técnicas é conhecida como *Descoberta de Conhecimento Apoiado por Computador (Knowledge Discovery - KD)*. Existem duas abordagens dentro dessa área chamadas de *Descoberta de Conhecimento em Dados Estruturados (Knowledge Discovery in Databases - KDD)* e *Descoberta de Conhecimento em Dados não Estruturados (Knowledge Discovery from Text - KDT)*, onde essa segunda abordagem, *KDT*, tem uma forte relação com a mineração de texto por depender de seus resultados para complementar outros processos como a clusterização, categorização, análises textuais entre outras [3].

III. Processo e Técnicas de Mineração de Texto

Para extrair o conhecimento a partir da base de dados com textos e informações não estruturadas, é preciso seguir um processo com algumas etapas e utilizar de técnicas para conseguir chegar ao resultado.

As etapas de processamento são dividas da seguinte forma: coleta, pré-processamento, indexação, mineração e análise [1]. Cada uma delas está melhor explicada logo abaixo:

- **Coleta:** Constitui na formação da base de documentos podendo ser através de informações de bases de histórico, robôs de *crawling* ou outras fontes de dados que façam parte do contexto que será explorado.
- **Pré-Processamento:** Preparação dos dados através da aplicação de técnicas como o *Processamento de Linguagem Natural (Natural Language Processing - NLP)* [4].
- **Indexação:** É feito uma preparação através de palavras chaves para ter um melhor desempenho na busca dos documentos armazenados.
- **Mineração:** Processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis disponíveis nos dados [5].
- **Análise:** Análise humana para validar se o conhecimento extraído trouxe o valor esperado.

Essa é uma das abordagens possíveis para a mineração de texto. Outros processos podem se tornar mais complexos ou mais refinadas dependendo da base de dados que será analisado e do resultado esperado. Algumas outras aplicações podem ser vistas em [3]–[5].

IV. Aplicações Práticas

A aplicação das técnicas de mineração de dados e mineração de textos são de grandes utilidades para diversas áreas. Havendo a disponibilidade da base de dados para ser analisado, o processo pode ajudar a trazer o conhecimento que o cliente esta buscando e auxiliar na tomada de decisões estratégicas ou na análise dos seus resultados.

Como amostra do poder de alcance dessas técnicas, veremos algumas aplicações práticas que mostram como elas foram utilizadas por áreas distintas para conseguirem se beneficiar dos resultados obtidos.

A. Industria

O sistema de solução de problemas *CASSIOPEE*, desenvolvido como parte de uma parceria entre a empresa americana *General Electric* e a francesa *SNECMA (Société Nationale d'Étude et de Construction de Moteurs d'Aviation)* foi aplicado por três grandes companhias aéreas europeias para diagnosticar e prever problemas para o Boeing 737, resultando num *cluster* com famílias de falhas.

O sistema recebeu o prêmio de inovação *Manago and Auriol* em 1996 [5].

B. Educação

Três pesquisadores da Universidade de *Limerick*, na Irlanda, utilizaram a base de dados de uma media social muito conhecida de perguntas e respostas voltado para programação de computadores, o *StackOverflow*, para tentarem identificar as maiores dificuldades e desafios encontrados pelos programadores.

Os resultados foram diversos agrupamentos com uma gama enorme de informações sobre as perguntas e respostas analisadas. Para citar algumas, foram listados artigos com maiores números de visualizações, termos mais buscados e categoria com maior número de citações.

Como alguns dos exemplos desse estudo, a categoria mais buscada foi *Software design patterns*, o título com mais frequência encontrada foi *Same-origin policy* e o termo mais pesquisado dentro da categoria de Sistemas Operacionais foi *regular expressions* [6].

Esse estudo pode muito bem ser aplicado a qualquer outra base de dados do tipo perguntas e respostas, analisando informações sobre outras disciplinas que estudantes de outra áreas utilizam, mapeando suas maiores dificuldades e trabalhando em possíveis soluções e melhoria na educação.

C. Inteligência Competitiva

Esse caso mostra como foram utilizadas as técnicas de *KDT* e *KDD* para empresas conseguirem uma abordagem pró-ativa em relação ao mercado aplicado a Inteligência Competitiva.

Com base em informações internas das companhias e dados abertos dos concorrentes, foi possível mapear características que ajudaram a empresa na tomada de decisão importantes, fazendo ela se posicionar mais estrategicamente em relação aos competidores.

Para isso, a utilização das duas técnicas foram indispensáveis, já que os dados analisados foram de um volume muito grande e com crescimento constante, tendo que ser atualizados e reavaliados com certa frequência [7].

D. Redes Sociais

As redes sociais são os grandes responsáveis pelo crescimento da quantidade de informações disponíveis nos dias atuais. A partir de seus dados, é possível retirar informações valiosas que ajudam empresas em seus negócios. Desde a venda de produtos, comportamento de consumo, perfil social e até mesmo o que você come, ou vai comer pode ser analisado através das suas redes sociais.

Seguindo essa tendência, três pesquisadores de universidades americanas analisaram dados do Twitter e do Facebook das três maiores empresas de Pizza dos Estados Unidos: Pizza Hut, Domino's Pizza e Papa John's Pizza [8].

O estudo mostrou como as medias sociais podem influenciar na tomada de decisão dos clientes de escolher entre uma empresa e outra. Toda essa análise foi feita através dos dados não estruturados como os *posts*, *likes*, comentários e seguidores para poder abordar o usuário da melhor maneira possível e transformando a capitação em resultado. Tudo isso através da coleta dos dados e utilização da mineração de texto.

Referências

[1] C. Aranha and E. Passos, "A tecnologia de mineração de textos," *JRESI-Revista Elerônica de Sistemas de Informação - Lab.ICA Elétrica PUC-Rio*, no. 2, 2006.

[2] S. P. Nagarkar and R. Kumbhar, "Text mining - an analysis of research published under the subject category 'information science library science in web of science database during 1999-2013," *Emerald Insight*, vol. 64, no. 3, February 2015.

[3] E. A. M. Morais and A. P. L. Ambrósio, "Mineração de textos," *Instituto de Informática - Universidade Federal de Goiás*, no. 7, Dezembro 2007.

[4] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *EMERGING TECHNOLOGIES IN WEB INTELLIGENCE*, vol. 1, no. 1, August 2009.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, 1996.

[6] A. Joorabchi, M. English, and A. E. Mahdi, "Text mining stackoverflow - an insight into challenges and subject-related difficulties faced by computer science learners," *Emerald Insight*, vol. 29, no. 2, August 2015.

[7] S. Loh, L. K. Wives, and J. P. M. Oliveira, "Descoberta proativa de conhecimento em textos: Aplicações em inteligência competitiva," *PPGC-UFRGS*, vol. 69, 2002.

[8] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management*, no. 33, February 2013.