Mario Caires Pereira

Estudo de Técnicas de Mineração de Texto Aplicadas na Classificação de Artigos Cientificos

Projeto de Graduação em Computação submetido à Universidade Federal do ABC para a obtenção dos créditos na disciplina Projeto de Graduação em Computação I do curso de Ciência da Computação

Orientador: Prof Dr Thiago Ferreira Covões

Universidade Federal do ABC

22 de Agosto de 2019

Resumo

Milhares de trabalhos científicos são desenvolvidos todos os anos em universidades públicas e privadas, entre eles: iniciações cientificas, trabalhos de conclusão de curso e teses de mestrado ou doutorado. O atual método para analisar tais trabalhos, é por meio de bancas avaliadoras, que na maioria das vezes é formada pelo orientador e dois (ou mais) docentes convidados pelo mesmo. Este projeto visa utilizar técnicas de mineração de textos aplicadas na classificação de trabalhos científicos, por meio da leituras de dados não estruturados presentes na Web, a partir dessa classificação, pretende-se identificar as palavras mais relevantes do trabalho. Com o intuito de, gerar uma banca coerente ao assunto abordado no artigo, através da indicação de docentes que atuam na área.

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

Desde o surgimento dos sistemas computacionais, um dos principais objetivos de empresas e organizações têm sido o armazenamento de dados. Nas ultimas décadas, em virtude dos avanços tecnológicos, a capacidade desse armazenamento é cada vez maior. Devido a isto, no final da década de 80 surge o termo Mineração de Dados (MD), também conhecido como Descoberta do Conhecimento em Banco de Dados.

Mineração de Dados possui o objetivo de extrair informações e padrões por meio da análise de grandes quantidades de dados, que previamente eram incompreensíveis ou desconhecidos. É uma área interdisciplinar, que envolve banco de dados, inteligência artificial, aprendizado de máquinas, estatística, entre outras. Dentre as diferentes possibilidades de aplicações de MD, pode-se destacar: mercado financeiro, identificando segmentos de mercado; tomadas de decisão, filtrando informações relevantes; marketing, direcionando mensagens promocionais para um determinado público alvo [Camilo e Silva 2009].

A MD realiza o estudo de dados estruturados, ou seja, eles são organizados conforme a definição de uma rígida estrutura. Essa disposição geralmente é realizada via linhas e colunas, permitindo "etiquetar"os dados, como por exemplo: banco de dados, planilha eletrônicas, arquivo CSV. Porém em razão dos grandes avanços tecnológicos de hardware e software voltados para a Web, em especial as redes sociais, a criação de conteúdos de textos, áudio e

imagem aumentou significantemente. Esses conteúdos reapresentam dados não estruturados, isso porque não hà necessidade se preocupar com campos pré definidos, restrições e limites. O usuário pode mesclar tipo de dados, como texto e imagem, vídeo e áudio, ou seja, o oposto do que seria uma estrutura rígida de um dado estruturado.

Atualmente, mais de 80% do conteúdo digital gerado no mundo é do tipo não estruturado, gerando a necessidade do desenvolvimento técnicas capazes de transformar estes dados em dados estruturados. Por exemplo, no caso de textos este processo é conhecido como Mineração de Textos (MT) [Feldman e Sanger 2006].

Mineração de Textos, também conhecido como Descoberta de Conhecimento em Textos, utiliza técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras. Envolve a aplicação de algoritmos computacionais que processam textos e identificam informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta, pois a informação contida nestes textos não podem ser obtida de forma direta, uma vez que, em geral, estão armazenadas em formato não estruturado [Morais e Ambrósio 2007].

Há duas abordagens que podem ser utilizadas na mineração de textos, são elas a Análise Semântica e Análise Estatística. A primeira realiza a interpretação das palavras conforme um ser humano faria, ou seja, por meio do significado da palavra, o contexto na qual ela está inserida bem como conhecimentos morfológicos, sintáticos e semânticos. Na Análise Estatística, as palavras são valoradas mediante a frequência de sua aparição nos dados, não importando a sua contextualização. As abordagens podem ser aplicadas separadamente ou em conjunto, de acordo com a necessidade do problema [Junior 2007].

Entre as principais técnicas de Mineração de Textos, as tarefas de Agrupamento e Classificação recebem especial atenção na literatura [Aggarwal e Zhai 2012]. Agrupamento consiste em agrupar automaticamente os documentos em grupos de acordo com a sua similaridade. A Classificação, também chamada Categorização de Textos, é utilizada para classificar um conjunto de documentos em uma ou mais categorias (classes) pré-definidas [Passini 2012].

As aplicações da mineração de textos são variadas tanto nas áreas científicas quanto comerciais. Um grande exemplo é a utilização desta técnica na medicina, onde diariamente milhares de informações de textos são geradas (prontuários, registros hospitalares, receitas, fichas de pacientes), auxiliando médicos com diagnóstico de doenças e recomendação de tratamentos. Há um software chamado Medline, que utiliza dos conceitos apresentados, trabalhando como base de dados bibliográficos da Biblioteca Nacional de Medicina dos Estados Unidos [Pezzini 2017].

No meio de diversas aplicações, uma nova área de pesquisa está emergindo, a "Mineração de Dados Educacionais"("*Educational Data Mining*", ou EDM) que possui o objetivo de estudar dados coletados no âmbito educacional e responder questões como: métodos para melhorar a aprendizagem do estudante, desenvolver ambientes educacionais mais eficazes, identificar se um aluno está confuso ou desmotivado com o método de aprendizagem e assim realizar adequações para sanar essa deficiência [Baker, Isotani e Carvalho 2011].

Neste projeto a mineração de textos será aplicada no ambiente acadêmico, com a finalidade de melhorar o método de como bancas avaliadoras são formadas atualmente. Possibilitando uma banca coesa com o tema apresentado no trabalho e consequentemente uma avaliação mais enriquecedora.

CAPÍTULO 2

JUSTIFICATIVA

A ideia do projeto surgiu após a indagação de como seria possível aprimorar o método como bancas avaliadoras são formadas nas universidades, com o intuito de automatizar o processo e ao mesmo tempo, indicar docentes que de fato trabalhem na área abordada pelo trabalho. Isso porque, a função da banca não é apenas a de atribuir uma nota, mas também de propor melhorias para o desenvolvimento do trabalho. Essa contribuição é efetiva quando a mesa é composta por profissionais que dominem o tema do projeto.

Atualmente o processo para construção de uma banca é engessado, no qual o orientador do trabalho é responsável por convidar docentes a sua escolha para compor a mesa avaliadora. Porém não é necessário que estes docentes possuam expertise na área de pesquisa apresentada no trabalho, nestes casos, a banca não possuirá uma colaboração positiva para o projeto, já que a análise será superficial, devido a falta de familiaridade com o assunto. Além disso, o processo é burocrático e na maioria das vezes o orientador convida docentes com os quais ele possui maior afinidade, não levando em conta o ramo de atuação do mesmo.

Com a aplicação de técnicas de mineração de textos, é possível identificar os docentes mais apropriados para qualificar o trabalho em questão de maneira assertiva, evitando assim, a seleção daqueles que não possuem proximidade ao tema abordado. Somado a isso, o

processo seria automatizado, retirando a função do orientador de selecionar os docentes da mesa, impossibilitando que critérios pessoais sejam relevantes na decisão.

CAPÍTULO 3

OBJETIVOS

Será desenvolvido um processo no qual, técnicas de mineração de textos serão aplicadas, afim de identificar os docentes que possuem maior afinidade com a área de pesquisa do trabalho sujeito a avaliação. Para tanto, serão utilizados o Currículo *Lattes*[1] dos docentes, com o intuito de identificar suas principais áreas de pesquisas.

Como objetivo secundário, visa-se realizar um agrupamento de docentes baseado em suas áreas de pesquisas. A ideia é semelhante ao objetivo principal, porém ao invés da leitura de um trabalho, serão analisados os currículos dos docentes e assim associar aqueles possuem maiores semelhanças. Os grupos serão de extrema utilidade, pois atualmente nos sites da UFABC as informações sobre os docentes e suas respectivas áreas de atuações, em sua maioria estão incorretas e/ou desatualizadas. Devido a isso, os discentes acabam enfrentando diversas dificuldades entre elas, a escolha de um professor orientador que atue na área de pesquisa do projeto ou identificar um docente apto a sanar dúvidas referentes a uma determinada área.

[1] A Plataforma Lattes é uma plataforma virtual criada e mantida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela qual integra as bases de dados de currículos, grupos de pesquisa e instituições, em um único sistema de informações, das áreas de Ciência e Tecnologia, atuando no Brasil.

CAPÍTULO 4

METODOLOGIA

Dado um trabalho científico como parâmetro de entrada, o processo realizará a leitura de todas as palavras presentes no texto, armazenando-as para posterior consulta. Feito isso, serão acessados os currículos *Lattes* dos docentes, identificando as palavras presentes em sua formação acadêmica, área de atuação, projetos de pesquisa, prêmios recebidos, produções realizadas, bancas e orientações. As palavras encontradas na *Plataforma Lattes* serão comparadas com as do trabalho, identificando aquelas que são semelhantes. Docentes que obtiverem maior correspondência de palavras, consequentemente são aqueles que atuam em áreas relacionadas à apresentada no trabalho.

Sabendo que o processo de Mineração de Textos é representado por cinco grandes etapas: identificação do problema, pré-processamento, extração de padrões, pós processamento e utilização do conhecimento, as próximas secções possuem a finalidade de identificar essas etapas no problema que será estudado [Rezende, Marcacini e Moura 2011]

4.1 Identificação do Problema

Utilizar técnicas de mineração de textos na recomendação dos docentes mais apropriados para comporem a banca avaliadora de um determinado trabalho. Para isso, serão utilizado

as informações presentes na *Plataforma Lattes* dos docentes, tais como seus projetos de pesquisas, orientações, trabalhos realizados, entre outras.

4.2 Pré-Processamento

O desafio inicial é verificar que parte da *Plataforma Lattes* contém os dados que deverão ser analisadas pelo processo. Posteriormente, identificar um método ou ferramenta que realize a leitura desse conteúdo e disponibilize para uso.

Para que o processo realizar as comparações entre as palavras, é de extrema importância que o texto armazenado esteja normalizado. Assim, é preciso aplicar algumas técnicas para "limpar"o texto, tais como *stopwords* e *stemming*, eliminando redundâncias e/ou variações morfológicas. Após a normalização do texto, preposições, plurais, letras maiúsculas e acentuações não irão interferir na análise do processo.

4.3 Extração de Padrões

Aqui deverá ser aplicado uma técnica ou modelo matemático para identificar entre as palavras armazenadas, aquelas que serão *key words* (palavras chave) e assim, serem utilizadas nesta etapa. Para cada palavra chave do texto de entrada, será realizada uma comparação entre as palavras chave retiradas da *Plataforma Lattes* dos docentes. Isso posto, os docentes que obtiverem um maior número de correlação, são aqueles que possuem maior familiaridade com o assunto abordado no trabalho e portanto os mais indicados para preencherem a banca. Visando o objetivo secundário, o processo será similar, porém as palavras chave comparadas serão somente aquelas identificadas na *Plataforma Lattes*, e o objetivo da correlação será o agrupamento daqueles que atuam na mesma área.

4.4 Pós Processamento

Tendo em vista a informação extraída, a formação de bancas coerentes com o tema abordado será realizada de maneira simples, pois bastará selecionar os docentes que foram identificados com maior familiaridade ao assunto. Além disso, as informações dos docentes são coletadas de uma página *web*, elas estarão atualizadas de acordo com as publicações e participações em pesquisas de cada um.

CAPÍTULO 5

FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo serão discutidos conceitos e técnicas referentes a Mineração de Textos. As etapas da MT apresentadas no Capítulo 4 serão utilizadas para identificar em qual momento do processo cada técnica é aplicada.

5.1 Pré-Processamento

O objetivo dessa etapa consiste em transformar o conjunto de documentos em uma base mais limpa, na qual o trabalho de representação, processamento dos dados e a consequente interpretação destes, possam ser realizadas de maneira mais rápida e eficiente [Passini 2012]. Segue a descrição de algumas das técnicas usualmente utilizadas na preparação dos dados:

5.1.1 Identificação de Termos

5.1.1.1 Identificação de Termos Simples

É aplicado um analisador léxico para identificar as palavras presentes no documento, eliminando símbolos e caracteres indesejados, tais como hífen e vírgula. Nesta etapa os termos podem ser convertidos para letras minusculas ou maiúsculas e tabulações convertidas a es-

paços simples, adequando os temos de acordo com a objetivo da análise [Morais e Ambrósio 2007].

5.1.1.2 Identificação de Termos Compostos

Há diversos termos que possuem diferentes significados quando descritos por meio da utilização de duas ou mais palavras adjacentes, são conhecidos como *Word-phrase formation*, podemos citar como exemplo o termo "Inteligência Artificial". Uma maneira de reconhecer estas palavras, consiste em identificar os termos que co-ocorrem com a maior frequência no documento, e posteriormente validar ou não as expressões [Morais e Ambrósio 2007].

5.1.2 Stopwords

No processo de análise de textos, é necessário identificar palavras que não demonstram relevância, possibilitando assim a sua remoção. Pode-se citar como exemplo os artigos, preposições, pronomes, advérbios e outras classes de palavras auxiliares. Estes termos formam a maior parte dos textos da língua portuguesa, não agregando valor ao entendimento do texto analisado [Passini 2012].

As *stopwords* formam um "dicionário negativo", também conhecido como *stoplist*. Assim, ao realizar a análise de um texto, as palavras encontradas no dicionário são identificadas como *stopwords*, resultando na remoção dos termos [Morais e Ambrósio 2007].

5.1.3 Stemming

Stemming é uma técnica de redução de termos a um radical comum, a partir da análise das características gramaticais dos elementos, como grau, número, gênero e desinência. Tem o objetivo de retirar os sufixos e prefixos das palavras, e encontrar a sua forma primitiva. Assim, as palavras no plural ou derivadas são reduzidas a um radical único, simplificando a representação dos termos envolvidos no documento [Passini 2012].

Dois erros típicos que costumam ocorrer durante o processo de *stemming* são *overstemming* e *understemming*. *Overstemming* ocorre quando não só o sufixo, mas também parte

do radical é retirado da palavra. Já *understemming* ocorre quando o sufixo não é removido, ou é apenas removido parcialmente [Uber 2004].

Há diversos algoritmos de *Stemming* desenvolvidos, porém eles devem ser projetados para o processamento de um idioma em específico. Como neste projeto serão analisados textos na língua portuguesa, destacam-se três algoritmos: a versão para português do algoritmo de PORTER [**?**], o Removedor de Sufixo da língua Portuguesa (RSLP), proposto por Orengo e Huyck [**?**] e o algoritmo STEMBR, proposto por Alvares [**?**] [Passini 2012].

Foram realizados estudos para identificar a técnica mais eficiente, comparando o desempenho dos três algoritmos citados anteriormente. O RSLP foi considerado o mais eficiente devido a menor taxa de erros de *overstemming* e *understemming* [**?**].

5.2 Extração de Padrões

Essa é a principal etapa do processo de Mineração de Textos, nela ocorre a busca efetiva por conhecimentos inovadores e úteis a partir dos dados textuais. A aplicação dos algoritmos, fundamentados em técnicas que procuram, segundo determinados paradigmas, visa explorar os dados de forma a produzir modelos de conhecimento.

5.2.1 Keywords

As palavras mais frequente em um texto (com exceção das *stopwords*) geralmente possuem um maior significado para o entendimento do assunto abordado no documento. Há duas maneiras de calcular a relevância destas palavras, a primeira delas é por meio da frequência que ela aparece no texto ou por meio da sua posição sintática.

Neste projeto será utilizado a análise baseada na frequência, mediante a atribuição de um "peso"para cada palavra. Há diversas maneiras de realizar o cálculo deste peso, a seguir serão descritos três métodos: frequência absoluta, frequência relativa e frequência inversa de documentos [Morais e Ambrósio 2007].

5.2.1.1 Frequência Absoluta

Também conhecido como *term frequency*(TF) é a técnica mais simples de se calcular o peso de uma palavra. Basta contabilizar a quantidade de vezes que o termo aparece no documento. Porém, por não levar em conta o tamanho do documento, palavras pouco frequentes em um pequeno texto, podem ter o mesmo peso que palavras muito frequentes em grandes documentos [Morais e Ambrósio 2007].

5.2.1.2 Frequência Relativa

Segundo Santos [Wives 1999] esta é a técnica mais comum para a identificação do quanto uma determinada palavra é importante para um documento, de acordo com o número de ocorrências desta palavra no mesmo. Segue a fórmula da frequência relativa:

$$F_{relX} = \frac{F_{absX}}{N} \tag{5.1}$$

Onde:

F_{rel}: frequência relativa de uma palavra x em um documento;

F_{abs}: número de vezes que a palavra aparece no documento;

N: número total de palavras no documento;

5.2.1.3 Frequência Inversa de Documentos

Essa técnica leva em conta a quantidade de documentos nos quais o termo aparece, somado a frequência absoluta dos termos. Assim, as palavras que aparecem em poucos documentos têm sua importância aumentada, pois geralmente são as mais discriminantes [Morais e Ambrósio 2007]. A fórmula para o cálculo da frequência inversa é:

$$Peso_{td} = \frac{Freq_{td}}{DocFreq_{td}} \tag{5.2}$$

Onde:

$Peso_{td}$: grau de relação entre o termo t e o documento d;

$Freq_{td}$: número de vezes que o termo t aparece no documento d;

$DocFreq_{td}$: número de documentos que o termo t aparece;

CAPÍTULO 6

CRONOGRAMA

Esta secção possui o objetivo de apresentar o cronograma deste projeto. A Tabela 6.1 contém uma visão geral das atividades descritas abaixo:

(A) - Levantamento Bibliográfico na Área de Mineração de Dados e Textos - Pesquisar livros e artigos que serão utilizados para o embasamento teórico do projeto;

(B) - Estudo de Técnicas de Mineração de Texto - Identificar qual a melhor técnica a ser utilizada na implementação do processo desejado;

(C) - Implementação do Processo - Etapa na qual a programação do processo ocorrerá;

(D) - Análise de Resultados- Identificar se os resultados alcançados foram os esperados, bem como discutir futuras propostas para melhorias;

(E) - Escrita do Relatório Técnico - O desenvolvimento do texto será realizado durante todo o ano, de acordo com o andamento do projeto.

Tabela 6.1: Cronograma

Atividades	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
A	X	X	X								
B				X	X	X	X				
C					X	X	X	X	X		
D								X	X	X	X
E	X	X	X	X	X	X	X	X	X	X	X

REFERÊNCIAS

[Aggarwal e Zhai 2012]AGGARWAL, C. C.; ZHAI, C. *Mining text data.* [S.l.]: Springer Science Business Media, 2012.

[Baker, Isotani e Carvalho 2011]BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. *Brazilian Journal of Computers in Education*, v. 19, n. 02, p. 03, 2011.

[Camilo e Silva 2009]CAMILO, C. O.; SILVA, J. C. da. *Mineraçãode Dados: Conceitos, Tarefas, Métodos e Ferramentas.* [S.l.], 2009.

[Feldman e Sanger 2006]FELDMAN, R.; SANGER, J. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* [S.l.: s.n.], 2006.

[Junior 2007]JUNIOR, J. R. C. Desenvolvimento de uma metodologia para mineração de textos. *Departamento de Engenharia Elétrica, Pontífica Universidade Católica do Rio de Janeiro*, 2007.

[Morais e Ambrósio 2007]MORAIS, E. A. M.; AMBRÓSIO, A. P. L. *Mineração de Textos.* [S.l.], 2007.

[Passini 2012]PASSINI, M. L. C. *MINERAÇÃO DE TEXTOS PARA ORGANIZAÇÃO DE DOCUMENTOS EM CENTRAIS DE ATENDIMENTO*. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, 2012.

[Pezzini 2017]PEZZINI, A. Mineração de textos: Conceito, processo e aplicações. *REAVI-Revista Eletrônica do Alto Vale do Itajaí*, v. 5, n. 8, p. 58–61, 2017.

[Rezende, Marcacini e Moura 2011]REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. *Revista de Sistemas de Informação da FSMA*, v. 7, p. 7–21, 2011.

[Uber 2004]UBER, J. L. Descoberta de conhecimento com o uso de text mining aplicada ao sac. Universidade Regional de Blumenau, 2004.

[Wives 1999]WIVES, L. K. Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de"clustering". 1999.