Universidade Federal de Santa Catarina

Text Mining

Data Mining — INE5644

Augusto Fredigo Hack
Luis Felipe Nunes
Matheus Hoffmann Silva
Thiago Thalison Firmino de Lima

Florianópolis, 19 de outrubro de 2013

Conteúdo

1 Introdução

Com a popularização da internet e a consolidação dos mecanismos de busca, encontramos hoje uma grande quantidade de dados, algo em torno de 1,8 zettabytes ou 1,8 Trilhões de GB a maioria desses dados encontra-se de maneira desestruturada ou semi estruturada, muitos desses dados está armazenada na forma de texto como e-mails, artigos e documentos digitalizados de uma forma geral, dado esse cenário , este trabalho se propõe a mostrar as técnicas desenvolvidas até a presente data para Descoberta de Conhecimento em Textos.

Palavras Chaves: Text Mining, Data Mining, Banco de Dados.

2 Text Mining *vs.* Data Mining

O Data Mining é o processo de exploração de grande quantidade de dados com o intuito de se descobrir padrões para formulação de um conhecimento até então não exposto de maneira explícita.

O Text Mining, inspirado no Data Mining, refere-se ao processo de Descoberta de Conhecimento em Texto conhecida pela sigla em inglês como KDT(Knowledge Discovery in Texts), consiste na obtenção de informação a partir de texto em linguagem natural ou passível de interpretação, o Text Mining extrai informação de dados estruturados ou semi estruturados, enquanto o Data Mining extrai informação de dados estruturados.

3 Técnicas de Pré-processamento de Textos

A preparação dos dados é a primeira etapa do processo de Text Mining, e, envolve a seleção dos dados que constituem a base de textos de interesse e o trabalho inicial para tentar selecionar o núcleo que melhor expressa o conteúdo destes textos. O objetivo desta etapa é tentar identificar similaridades em função da morfologia ou do significado dos termos nos textos, bem como prover uma redução dimensional do texto analisado [2].

Nesta seção serão apresentadas algumas das técnicas de pré-processamento de textos, sendo elas: Filtering, Tokenization, Stemming, Stopword removal, Pruning, Thesaurus, Cálculo de Relevância de Palavra, Vetorização de Textos.

3.1 Filtering

A primeira etapa para do processamento é a remoção dos caracteres de pontuação, esse caracteres de maneira geral não alteram o significado do documento, apesar da remoção da pontuação poder alterar o significado das frases, no geral o significa do documento fica inalterado, o viés do documento é definido pelas palavras mais significantes no texto e os caracteres de pontução não tem valor para o seu processamento.

3.2 Tokenization

Depois de removida a pontuação do documento, o processamento de linguagem natural é usado para identificar as palavras das frases e pode ser usado para tentar identificar o significado das frases, tendo a estrutura das frases o seu significado pode ser expresso em termos

das relações de suas palavras. Durante o processo as palavras usadas no corpo do texto são classificadas para uso nas etapas a seguir.

3.3 Stemming

Tendo as palavras separadas e classificadas, elas são reduziadas para sua forma mais básica, para seu radicais, a fim de remover estilos de escrita entre documentos e expressar o significado dos documentos de maneira consistênte que os deixem comparáveis.

3.4 Stopword Removal

Textos tem naturalmente um grande volume de dados, o tempo de análise de vários corpos de texto cresce rapidamente, para amenizar o tempo de processamento é necessário diminuir do total de dados que será processado dados que não tem grande valor. Palavras que são muito utilizadas no texto e que não trazem significado direto a uma frase são removidas, tanto para diminuir o corpo do texto e para limitar o conjunto às que trazem mais significado ao corpo de texto, assim as palavras muito frequêntes na escrita são consideradas como de pouco valo e na etapa de stopword removal são removidas, tudo aquilo que não é substantivo, adjetivo ou verbo é removido.

3.5 Pruning

Mesmo depois da etapa de 'Stopword removal', ainda existem palavras que não trazem valor ao texto, por serem usadas com muita frequência ou pouca frequência. A ideia sendo que palavras muito frequêntes em um corpo de texto também será frequênte nos outros corpos de texto, não trazendo valor as analises, e palavras que pouco aparecem têem uma chance pequena de reaparecer em outros textos, portanto na etapa de pruning as palavras que aparecem com muita e pouca frequência são removidas do texto.

3.6 Thesaurus

Por fim, como palavras diferentes podem ter o mesmo significado, são normalizadas, nessa etapa um dicionários de thesaurus, que relacionam palavras com o mesmo significado, são usados para normalizar o texto em um corpo bem padronizado.

3.7 Cálculo de Relevância de Palavra

Nem todas as palavras presentes em um documento possuem a mesma importância. Os termos mais freqüentemente utilizados (com exceção das stopwords) costumam ter significado mais importante, assim como as palavras constantes em títulos ou em outras estruturas, uma vez que provavelmente foram colocadas lá por serem consideradas relevantes ou descritivas para a idéia do documento.

A idéia do cálculo da relevância de uma palavra dentro de um documento objetiva obter um peso referente ao uso do termo dentro do texto. Existem várias fórmulas para cálculo do peso. As mais comuns são baseadas em cálculos simples de frequência: frequência absoluta, frequência relativa, frequência inversa de documentos.

3.7.1 Frequência Absoluta

Também conhecida por freqüência do termo ou term frequency (TF), representa a medida da quantidade de vezes que um termo aparece em um documento. Essa é a medida de peso mais simples que existe, mas não é aconselhada em alguns casos, porque, em análise de coleções de documentos, não é capaz de fazer distinção entre os termos que aparecem em poucos ou em muitos documentos. Este tipo de análise também não leva em conta a quantidade de palavras existentes em um documento. Com isso, uma palavra pouco freqüente em um documento pequeno pode ter a mesma importância de uma palavra muito freqüente de um documento grande.

3.7.2 Frequência Relativa

Este tipo de análise leva em conta o tamanho do documento (quantidade de palavras que ele possui) e normaliza os pesos de acordo com essa informação. A frequência relativa (Frel) de uma palavra x em um documento qualquer é calculada dividindo-se sua frequência absoluta (Fabs) pelo número total de palavras no mesmo documento (N):

$$F_{rel}(x) = \frac{F_{abs}(x)}{N}$$

Figura 1.

3.7.3 Frequência Inversa de Documentos

A frequencia inversa de documentos busca normalizar termos frequentes com base na sua ocorrência em todos os documentos analisados. O cálculo da freqüência inversa de documentos (inverse document frequency - IDF) é realizado com base na informação da freqüência absoluta do termo no documento e da freqüência do termo em todos os documentos, e isto é capaz de aumentar a importância de termos que aparecem em poucos documentos, e, diminuir a importância de termos que aparecem em muitos, justamente pelo fato dos termos de baixa freqüência serem, em geral, mais discriminantes.

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

Figura 2.

3.8 Vetorização de Textos

A vetorização de textos consiste na representação de um texto na forma de um vetor de termos. A forma mais comum de vetorização de textos é associar cada termo com uma frequencia, onde cada documento é representado por um vetor de termos e cada termo possui um valor associado que indica o grau de importância (denominado peso) desse no documento. Portanto,

cada documento possui um vetor associado que é constituído por pares de elementos na forma (palavra 1, peso 1), (palavra 2, peso 2)...(palavra n, peso n).

Nesse vetor são representadas todas as palavras da coleção de documentos e não somente aquelas presentes no documento. Os termos que o documento não contém recebem grau de importância zero e os outros são calculados através de uma fórmula de identificação de importância. Isso faz com que os pesos próximos de um (1) indiquem termos extremamente importantes e pesos próximos de zero (0) caracterizem termos completamente irrelevantes (em alguns casos a faixa pode variar entre -1 e 1).

O peso de um termo em um documento pode ser calculado de diversas formas. Esses métodos de cálculo de peso geralmente se baseiam na contagem do número de ocorrências dos seus termos (freqüência), que foram explicados na seção 3.7.

4 Aplicações de Text Mining

Nesta seção serão apresentadas algumas das técnicas de text mining, sendo elas: Classificação e Clusterização de Documentos, Extração de Informações e Processamento de Linguagem Natural.

4.1 Extração de Informações em Textos

As técnicas de Data Mining tradicionais exigem que os dados de entrada, que serão 'minerados', já estejam em formato estruturado. Infelizmente, para muitas aplicações, as informações presentes em textos eletrônicos estão apenas disponíveis em formato livre e escritos em linguagem natural, chamados de não estruturados. Esta situação impossibilita a aplicação direta de Data Mining.

A técnica de Extração de Informação (Information Extraction) tem por objetivo localizar ítens específicos dentro de um documento textual não estruturado, para então estruturá-los em formatos processáveis por máquinas, por exemplo banco de dados relacional ou arquivos XML.

Assim como descrito na Figura X, o processo de Extração de Informação é requisito para que se aplique o Data Mining:

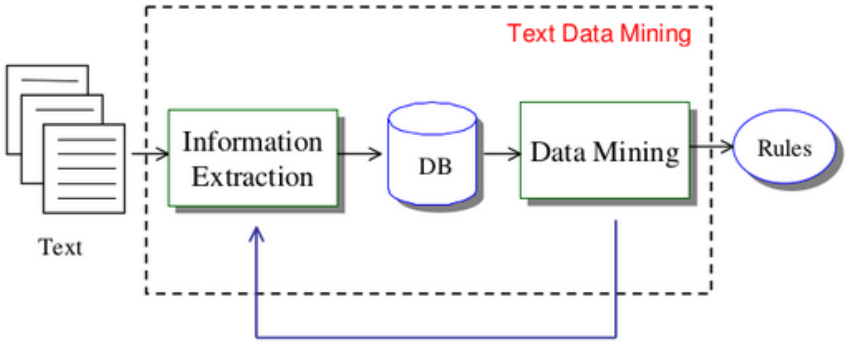

Figura 3 - Processo de Descoberda de Conhecimento por Text Mining.

A construção de sistemas de Extração de Informação automáticos é uma tarefa difícil. Um meio é a etiquetação manual de um pequeno conjunto de documentos, gerando um template que pode ser usado como entrada para um sistema IE. Este sistema pode então ser aplicado a conjuntos maiores de dados.

Pode-se exemplificar um sistema de IE simples: o sistema X tem o objetivo de adquirir informações sobre vagas de emprego. O Sistema X utiliza templates para estruturar dados. A Figura 4 exemplifica um documento a ser processado, que é trecho de um anúncio de vaga de emprego.

```
Title: Web Development Engineer
Location: Beaverton, Oregon

This individual is responsible for design and implementation
of the web-interfacing components of the AccessBase server,
and general back-end development duties.

A successful candidate should have experience that includes:

   One or more of: Solaris, Linux, IBM AIX, plus Windows/NT
   Programming in C/C++, Java
   Database access and integration: Oracle, ODBC
   CGI and scripting: one or more of Javascript,
                      VBScript, Perl, PHP, ASP

Exposure to the following is a plus: JDBC, Flash/Shockwave,
FrontPage and/or Cold Fusion.

A BSCS and 2+ years experience (or equivalent) is required.
```

Figura 4 - Exemplo de anúncio de emprego.

Então, à partir do processamento do texto pelo Sistema de IE X, obtêm-se um template preenchido contendo a informação extraída do documento de exemplo da Figura 4.

title:	“Web Development Engineer”
location:	“Beaverton, Oregon”
languages:	“C/C++”, “Java”, “Javascript”, “VBScript”, “Perl”, “PHP”, “ASP”
platforms:	“Solaris”, “Linux”, “IBM AIX”, “Windows/NT”
applications:	“Oracle”, “ODBC”, “JDBC”, “Flash/Shockwave”, “FrontPage”, “Cold Fusion”
areas:	“Database”, “CGI”, “scripting”
degree required:	“BSCS”
years of experience:	“2+ years”

Figura 5 - Template preenchido, como resultado do sistema exemplo de IE.

Após a construção de um sistema de IE que é capaz de extrair a informação desejada para uma determinada aplicação, os dados podem ser estruturados, eg. banco de dados. Técnicas de Data Mining podem então ser aplicadas para que finalmente seja possível gerar conhecimento à partir das informações coletadas.

4.2 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é a área da Ciência da Computação que tem por objetivo estudar mecanismos de processamento da linguagem falada e escrita, com o intuito de convertê-la para uma representação mais formal, e assim torná-la manipulável por programas de computador.

Segundo a Comissão Especial de Processamento de Linguagem Natural da Sociedade Brasileira de Computação - SBC (http://www.nilc.icmc.usp.br/cepln/, Acesso em: 06/10/2013), ‘A área de Processamento da Linguagem Natural (PLN), também denominada Linguística Computacional ou, ainda, Processamento de Línguas Naturais, lida com problemas relacionados à automação da interpretação e da geração da língua humana em aplicações como Tradução Automática, Sumarização Automática de Textos, Ferramentas de Auxílio à Escrita, Perguntas e Respostas, Categorização Textual, Recuperação e Extração de Informação, entre muitas outras...’.

O PLN é formalmente divido em 4 etapas, que são as analises morfológica, sintática, semântica e pragmática do texto. Não necessariamente deve haver uma ordem na execução de cada etapa, bem como não há necessidade de aplicá-las juntas, pois isto irá depender do domínio de aplicação no qual o PLN será utilizado.

4.2.1 Análise Morfológica

O objetivo da análise morfológica do texto é identificar a classe morfológica de uma palavra (substantivo, verbo, pronome, etc), bem como de sua inflexão, seja ela nominal (ex. gênero) ou verbal (ex. pessoa).

Uma das técnicas utilizadas para este tipo de análise é a utilização de tabelas de afixos (sufixos e prefixos), que associa sufixos e prefixos com radicais de palavras. A Figura 6 apresenta

uma tabela de sufixos, que são associados à radicais de palavras. Este tipo de estrutura permitiria identificar, por exemplo, que a palavra 'bonezinho' em um texto, é uma derivação da palavra boné e esta no diminutivo, pois foi utilizado o sufixo '-zinho'.

sufixo	substantivo (radical)	palavra derivada
-zinho	boné	bonezinho
	botão	botãozinho
-zinha	árvore	árvorezinha
	flor	florzinha

Figura 6

Os sistemas que fazem este tipo de análise são conhecidos como Finite State Transducer [1]. Nestes sistemas a entrada geralmente é o texto plano e a saída é o mesmo texto com tags que indicam as características morfológicas de cada palavra. A Figura 7 apresenta a saída de um Finite State Transducer capaz de analizar palavras em português, disponibilizado pela universidade holandeza Syddansk Universitet [2]. O texto de entrada utilizado foi 'O rapaz pegou o seu carrão'.

```
o [o] <artd> <dem> DET M S
rapaz [rapaz] <Hbio> N M S
pegou [pegar] <vt> <vi> V PS 3S IND VFIN
o [o] <artd> <dem> DET M S
seu [seu] <poss 3S/P> <si> DET M S
carrão [carro] <DERS> N M S
.
```

Figura 7

Analisando a saída acima é possível perceber que o programa conseguiu identificar que a palavra carrão deriva da palavra carro, e que a derivaçao ocorreu por um processo de sufixação, que é indicado pela tag <DERS>.

4.2.2 Análise Sintática

O objetivo da análise sintática é estabelecer as relações formais entre as palavras de uma frase, baseado nas regras gramaticais da linguagem na qual a frase foi escrita. Análise sintática de uma palavra é geralmente realizada utilizando três técnicas, são elas: Etiquetação das palavras, Divisão da palavra em Sintágmas Nominais e Verbais, Identificação da Função Sintática da Palavra.

A Etiquetação de palavras é processo de análise morfológica de cada palavra do texto. Este processo é importante porque geralmente as gramáticas se baseiam na classificação morfológica das palavras para estabelecer as relações entre elas.

O processo de divisão de uma frase em sintagmas nominais e verbais consiste na divisão de uma frase em segmentos cujo núcleo (palavra principal) será um nome (sintagma nominal) ou um verbo (sintagma verbal). Um exemplo seria a frase 'O João tropeçou na pedra', onde a estrutura 'O João' seria considerada o sintagma nominal, e 'tropeçou na pedra' o sintagma verbal da frase.

O processo de identificação da função sintática da palavra é a fase mais complexa deste tipo de análise, e geralmene é realizado com a aplicação de Gramáticas Livres de Contextos (GLC). A saída normalmente é uma àrvore cujos nodos folha são as palavras do texto analisado, e os nodos pai são as 'etiquetas sintáticas' para a palavra.

Para exemplificar este tipo de análise, considere a frase 'João Viu Maria', e a seguinte GLC:

S => NP, SV

SV => V, NP

Figura 8

Onde, S - Expressão; NP - Nome Próprio; SV - Sintagma Verbal; V - Verbo. A saída da aplicação da gramática sobre o texto poderia ser uma àrvore com a seguinte estutura:

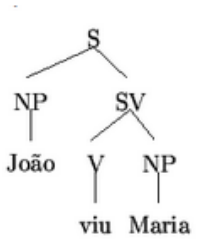

Figura 9

4.2.3 Análise Semântica

O processo de análise semântica do texto busca o mapeamento de sentenças de uma linguagem, visando a representação de seu significado, baseado nas construções obtidas nas análises morfológica e sintática.

Este tipo de análise é a mais difícil de ser implementada, pois envolve análise do significado de palavras e expressões, o que não é uma tarefa simples de ser realizada computacionalmente.

A técnica mais simples para realização da análise semântica é a contrução de regras semânticas utilizando gramáticas semânticas. A figura 10 mostra como seria a saída da aplicação da análise semântica em um texto. Nota-se que na anotação é identificado que a frase aborda uma ação de tropeçar, onde quem tropeça é o 'eu' (sujeito da frase), e quem é tropeçado é a 'pedra' (objeto indireto).

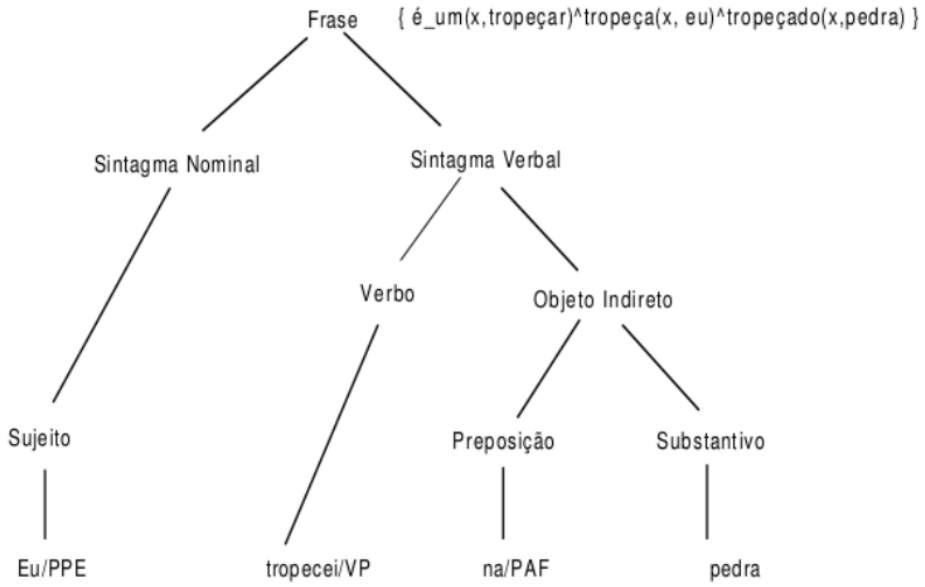

Figura 10

4.2.4 Análise Pragmática

A análise pragmática é processo de análise do significado de determinados termos do texto, baseado no contexto no qual ocorrem. Esta fase geralmente é necessária devido ao fato de que o texto geralmente é processado frase por frase. No entanto, é muito comum existirem frases que dependem de outra para que tenham o seu significado compreendido.

Uma das técnicas mais utilizadas neste fase é o algoritmo de Lappin e Leass [3] que consegue encontrar o substantivo referenciado por um pronome. Para isso ele cálcula a distância entre o pronome e os substântivos mais próximos, a distância mais próxima identifica um peso mais alto e distâncias maiores identificam pesos mais baixos. Além da distância o algoritmo também considera elementos de concordância como gênero e número.

5 Conclusão

Explorar e desenvolver sistemas que possam extrair informações de texto é um grande desafio, devido a dificuldade encontrada na análise de linguagens naturais e na quantidade de dados que precisa ser analisado para que se formule um conhecimento correto e que não deixe escapar dados essenciais, apesar disso, automatiar o processo de descoberta de conhecimento em texto, possui um grande potencial para auxiliar as organizações no processo de tomada de decisão.

6 Referências

[1] EBECKEN, N; LOPES, M; COSTA, M. *Mineração de Textos*, chapter 13, p. 337-370. Manole, 2003.

[2] LAPPIN, Shalom; Leass Herbert. *An Algorithm for Pronominal Anaphora Resolution*. Computational Linguistics 20, 535-561 (1994).

[3] MOONEY, Raymond J; NAHM, Un Yong. *Text Mining with Information Extraction*. Department of Computer Sciences, University of Texas, Austin, TX 78712-1188. 2005.

[4] Desconhecido. *tf–idf*. Disponível em: <http://en.wikipedia.org/wiki/Tf%E2%80%93idf>. Acesso em: 20/12/2013.

[5] Desconhecido. *Pré-processamento de Texto*. Disponível em: <http://pt.wikipedia.org/wiki/Pr%C3%A9-processamento_de_texto>. Acesso em: 20/12/2013.