PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

João Ribeiro Carrilho Junior

Desenvolvimento de uma Metodologia para Mineração de Textos

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica da PUC-Rio.

Orientador: Prof. Emmanuel Piseces Lopes Passos

Rio de Janeiro

Dezembro de 2007

Pontifícia Universidade Católica
do Rio de Janeiro

João Ribeiro Carrilho Junior

Desenvolvimento de uma Metodologia para Mineração de Textos

Dissertação apresentada como requisito parcial para obtenção do grau Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Dr. Emmanuel Piseces Lopes Passos
Orientador
Departamento de Engenharia Elétrica

Dra. Marley Maria Bernardes Rebuzzi Vellasco
Departamento de Engenharia Elétrica - PUC-RIO

Dr. Antonio Luz Furtado
Departamento de Informática - PUC-RIO

Dr. Christian Nunes Aranha
Cortex Intelligence

Dr. Ricardo Tanscheit
Departamento de Engenharia Elétrica - PUC-RIO

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico – PUC-Rio

Rio de Janeiro, 18 de dezembro de 2007

João Ribeiro Carrilho Junior

Graduou-se Bacharel em Informática pela PUC-Rio em 2004. Atua como analista de sistemas na Petrobras, principalmente no desenvolvimento de sistemas de apoio à decisão. Tem interesse na pesquisa de novos algoritmos, principalmente na área de Mineração de Textos.

Ficha Catalográfica

Carrilho Junior, João Ribeiro

Desenvolvimento de uma metodologia para mineração de textos / João Ribeiro Carrilho Junior ; orientador: Emmanuel Piseces Lopes Passos. – 2007.

96 f. ; 30 cm

Dissertação (Mestrado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Mineração de textos. 3. Dados não-estruturados. 4. Processamento de linguagem natural. 5. Aprendizado de máquina. 6. Recuperação de informação. I. Passos, Emmanuel Piseces Lopes. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

Agradecimentos

A Deus, por conceder-me mais esta benção em minha vida.

Aos meus pais, João e Tereza, e minha irmã, Fabiana, que tanto me apoiaram e me incentivaram.

À minha namorada, Beatriz, por ter acreditado em mim e estado ao meu lado em todos os momentos, sendo estes difíceis ou não.

Aos meus avós, Afonso e Sebastiana, que sempre me amaram e estiveram prontos a me ajudar.

Ao professor Emmanuel pelos seus valiosos ensinamentos e pela confiança depositada em meu trabalho.

Aos amigos Roberto e Fábio, que foram companheiros de estudos e que sempre estiveram prontos a colaborar.

Ao CNPq pelo apoio financeiro durante o primeiro ano do mestrado.

Resumo

Carrilho Junior, João Ribeiro; Passos, Emmanuel Piseces Lopes (Orientador). **Desenvolvimento de uma Metodologia para Mineração de Textos.** Rio de Janeiro, 2007. 96p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

A seguinte dissertação tem como objetivo explorar a Mineração de Textos através de um estudo amplo e completo do que atualmente é considerado "estado da arte". Esta nova área, considerada por muitos como uma evolução natural da Mineração de Dados, é bastante interdisciplinar e vem obtendo importantes colaborações de estudiosos e pesquisadores de diversas naturezas, como Lingüística, Computação, Estatística e Inteligência Artificial. Entretanto, muito se discute sobre como deve ser um processo completo de investigação textual, de forma a tirar máximo proveito das técnicas adotadas nas mais variadas abordagens. Desta forma, através de um encadeamento sistemático de procedimentos, pode-se chegar a uma conclusão do que seria a metodologia ideal para a Mineração de Textos, conforme já se chegou para a de Dados. O presente trabalho explora um modelo de processo, do início ao fim, que sugere as seguintes etapas: coleta de dados, pré-processamento textual, indexação, mineração e análise. Este sequenciamento é uma tendência encontrada em trabalhos recentes, sendo minuciosamente discutido nos capítulos desta dissertação. Finalmente, a fim de se obter enriquecimento prático, foi desenvolvido um sistema de Mineração de Textos que possibilitou a apresentação de resultados reais, obtidos a partir da aplicação de algoritmos em documentos de natureza geral.

Palavras-chave

Mineração de Textos; Dados Não-Estruturados; Processamento de Linguagem Natural; Aprendizado de Máquina; Recuperação de Informação.

Abstract

Carrilho Junior, João Ribeiro; Passos, Emmanuel Piseces Lopes (Advisor). **Development of a Methodology for text Mining.** Rio de Janeiro, 2007. 96p. MSc Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

The following essay is intended to explore the area of Text Mining, through an extensive and comprehensive study of what is currently considered "state of the art". This new area, considered by many as a natural evolution of the Data Mining, is quite interdisciplinary. Several scholars and researchers from fields like linguistics and computing, for instance, have contributed for its development. Nevertheless, much has been discussed on how complete dossier of textual investigation must be carried out, in order to take maximum advantage of the techniques adopted in various approaches. Thus, through a systematic sequence of procedures, one can come to a conclusion of what would be the ideal method for the Mining of documents, as one has come about Data. This work explores a model of process which suggests the following steps: collecting data, textual pre-processing, indexing, mining and analysis. This sequence is a tendency followed in some recent works and it is thoroughly discussed in the chapters to come. Finally, in order to obtain a practical enrichment, one developed a system of Mining of documents with which became possible the presentation of results, obtained from the application of algorithms in documents of a general nature.

Keywords

Text Mining; Unstructured Data; Natural Language Processing; Machine Learning; Information Retrieval.

Sumário

1 Introdução 12
1.1. Objetivos da Dissertação 13
1.2. Organização da Dissertação 13

2 Mineração de Textos: Fundamentos e Aplicações 14
2.1. Áreas de Conhecimento em Mineração de Textos 15
2.1.1. Processamento de Linguagem Natural 15
2.1.2. Ciência Cognitiva 16
2.1.3. Recuperação de Informação 16
2.1.4. Estatística 18
2.1.5. Aprendizado de Máquina 18
2.1.6. Inteligência Computacional 19
2.1.7. Mineração de Dados 19
2.1.8. *Web Mining* 20
2.2. Aplicações 21
2.2.1. Negócios 21
2.2.2. Direito 24
2.2.3. Medicina 24

3 Etapas da Metodologia de Mineração de Textos 26
3.1. Coleta 27
3.2. Pré-processamento 30
3.2.1. *Tokenization* (Atomização) 31
3.2.2. Correção Ortográfica 34
3.2.3. Redução do Léxico 35
3.2.4. Identificação do Início e Fim de Sentenças 42
3.2.5. Etiquetagem POS 45
3.2.6. Identificação de Entidades Nomeadas 45
3.2.7. *Parsing* (Análise Sintática) 46
3.3. Indexação 48
3.3.1. Representação de Documentos 48
3.3.2. Medidas de Similaridade entre Documentos 49
3.3.3. Listas Invertidas 51
3.3.4. Processamento de Consultas 52
3.3.5. Avaliação das Consultas 54
3.4. Mineração 56
3.5. Análise da Informação 56

4 Tarefas de Mineração de Textos 58
4.1. Categorização de Textos 58
4.1.1. Treinamento e Teste 60
4.1.2. Avaliação de Performance 62
4.1.3. *Naive Bayes* 64
4.2. *Clusterização* 67

4.3. Sumarização 69
4.4. Extração de Informação 71
4.5. Sistemas de Busca de Informação 73

5 Implementação e Estudo de Caso 74
5.1. Arquitetura Geral do Sistema 74
5.1.1. Módulo de Coleta 75
5.1.2. Módulo de Pré-processamento 79
5.1.3. Módulo de Indexação 81
5.1.4. Módulo de Mineração 82
5.1.5. Módulo de Análise de Resultados 84
5.2. Estudo de Caso: Identificação de Subjetividade em Pesquisas de Opinião 87

6 Conclusão 91

Referências bibliográficas 92

Lista de Figuras

Figura 1 – Componentes de um sistema de Recuperação de Informação 17

Figura 2 – Abordagens de *Web Mining* 21

Figura 3 - Diagrama que ilustra a metodologia de Mineração de Textos com o "encadeamento" de técnicas proposta por Aranha. 26

Figura 4 – "Linha de montagem" de um procedimento de *Tokenization.* 32

Figura 5 – Exemplo de um algoritmo de detecção de início e fim de sentenças... 44

Figura 6 - Árvore de Derivação simples para a frase "José comeu o bolo". 47

Figura 7 – Exemplificação do modelo "saco de palavras". 49

Figura 8 – Documentos "apontando" para seus *tokens.* 51

Figura 9 – Estrutura de Lista Invertida com os *tokens* "apontando" para os documentos. 52

Figura 10 - Classificação ternária de documentos. 59

Figura 11 – Utilização da estratégia *holdout* para treinamento e validação de classificadores. 61

Figura 12 –Validação Cruzada com 3-*folds*. 62

Figura 13 – Esquema básico da Tarefa de *Clusterização.* 67

Figura 14 – Agrupamento de não-hierárquico aglomerativo de documentos. 68

Figura 15 – Extração de Características de um documento. 71

Figura 16 – Diagrama Hierárquico de Funções do sistema implementado. 74

Figura 17 – Modelo MVC utilizado no desenvolvimento do sistema. As setas sólidas indicam associações diretas e as tracejadas indicam associações indiretas. 75

Figura 18 – Diagrama de classes parcial do sistema com as classes Corpus e Documento. 76

Figura 19 – Ciclo contínuo de execução da coleta na Internet através de *web crawlers.* 78

Figura 20 - Diagrama de classes contendo as novas classes introduzidas pela etapa de *tokenization* do módulo de pré-processamento..... 79

Figura 21 – Parte do diagrama de classes do sistema com a adição das partes envolvidas na indexação..... 81

Figura 22 – Diagrama de classes parcial do sistema com a inclusão das classes que compõem o módulo de mineração..... 83

Figura 23 – Gráfico de barras que mostra a relação entre faixas de freqüência e o número total de *tokens* presentes..... 86

Lista de Tabelas

Tabela 1 – As duas abordagens para a Análise de Textos e suas principais Áreas de Conhecimento 15
Tabela 2 – Principais focos no monitoramento de páginas na *web* 22
Tabela 3 – Resumo das principais coleções de texto usadas pela comunidade científica 29
Tabela 4 - Representação atributo-valor obtida à partir da etapa de Pré-processamento 30
Tabela 5- Exemplificação do resultado da execução de um subsistema de *Tokenization* que baseia-se em dicionários pré-estabelecidos e regras de formação. 33
Tabela 6 – Identificação e Remoção de *Stopwords* (os *tokens* descartados estão tachados) 39
Tabela 7 – *Stoplist* obtida automaticamente a partir de um sistema de Mineração de Texto pronto 39
Tabela 8 – Matriz de Confusão 85
Tabela 9 – Relatório de freqüência dos tokens nos corpus envolvidos no treinamento do classificador 86
Tabela 10 – Lista de stopwords padrão do Google. 88
Tabela 11 – Resultados dos experimentos realizados com o classificador Naive Bayes no problema de subjetividade versus objetividade. 89
Tabela 12 – Resultados dos experimentos realizados com o classificador SVM no problema de subjetividade versus objetividade. 89
Tabela 13 – Categorização obtida para novos exemplos utilizado o classificador *Naive Bayes.* 90