



Fábio de Azevedo Soares

**Mineração de Textos na Coleta Inteligente de Dados
na Web**

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.

Orientadores:

Prof. Emmanuel Piseces Lopes Passos

Profa.. Marley Maria Bernardes Rebuzzi Vellasco

Rio de Janeiro

Setembro de 2008



Fábio de Azevedo Soares

**Mineração de Textos na Coleta Inteligente de Dados
na Web**

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.

Profa. Marley Maria Bernardes Rebuzzi Vellasco
Orientadora

Departamento de Engenharia Elétrica – PUC-Rio

Prof. Emmanuel Piseces Lopes Passos
Co-Orientador

Departamento de Engenharia Elétrica – PUC-Rio

Prof. Antonio Luz Furtado

Departamento de Informática – PUC-Rio

Prof. Christian Nunes Aranha

Cortex Intelligence

Prof. José Eugenio Leal

Coordenador Setorial do Centro
Técnico Científico – PUC-Rio

Rio de Janeiro, 05 de setembro de 2008

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Fábio de Azevedo Soares

Graduou-se Bacharel em Ciência da Computação em 2006. Atua como analista de sistemas, principalmente no desenvolvimento de Sistemas de Apoio à Decisão. Tem interesse na pesquisa de novos algoritmos, principalmente, na área de Mineração de Textos e Aprendizado de Máquina.

Ficha Catalográfica

Soares, Fábio de Azevedo

Mineração de textos na coleta inteligente de dados na web / Fábio de Azevedo Soares; orientadores: Emmanuel Piseces Lopes Passos, Marley Maria Bernardes Rebuszi Vellasco. – 2008.

120 f. : il. ; 30 cm

Dissertação (Mestrado em Engenharia Elétrica)– Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Mineração de textos. 3. Coleta de dados. 4. Web crawling. 5. Recuperação da informação. I. Passos, Emmanuel Piseces Lopes. II. Marley Maria Bernardes Rebuszi Vellasco III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Aos meus pais,
Estevão e Miracema,
que não se casam de me ajudar na realização dos meus sonhos.

Agradecimentos

À Deus e a Todos Aqueles que me protegem, me iluminam e me guiam pelo caminho do bem, sem cujas permissões nada teria sido iniciado, quanto menos concluído.

Ao meu irmão, Tiago, que tanto me apoiou e me incentivou.

À minha namorada, Rakely, pela enorme paciência, imensa compreensão e apoio incondicional.

Ao professor Emmanuel Passos pela oportunidade de realizar este trabalho, pela confiança depositada em mim, pelo exemplo de fé e por toda lição de vida.

Ao meu amigo Cristian Klen por me apoiar, incentivar e ajudar na realização deste trabalho.

Ao meu amigo João Carrilho por ter muito bem me recebido quando aqui cheguei para trilhar este novo caminho.

Ao meu amigo Roberto Gomes pelo seu grande companheirismo e incentivo em todos os momentos.

Ao meu amigo Sérgio Ciglione pela amizade e companhia nos momentos de estudo.

Ao CNPq e CAPES pelo apoio financeiro.

Resumo

Soares, Fábio de Azevedo; Passos, Emmanuel Piseces Lopes (Orientador); Vellasco, Marley Maria Bernardes Rebuzzi (Orientadora). **Mineração de Textos na Coleta Inteligente de Dados na Web**. Rio de Janeiro, 2008. 120p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Esta dissertação apresenta um estudo sobre a utilização de Mineração de Textos no processo de coleta inteligente de dados na *Web*. O método mais comum de obtenção de dados na *Web* consiste na utilização de *web crawlers*. *Web crawlers* são *softwares* que, uma vez alimentados por um conjunto inicial de *URLs* (sementes), iniciam o procedimento metódico de visitar um *site*, armazená-lo em disco e extrair deste os *hyperlinks* que serão utilizados para as próximas visitas. Entretanto, buscar conteúdo desta forma na *Web* é uma tarefa exaustiva e custosa. Um processo de coleta inteligente de dados na *Web*, mais do que coletar e armazenar qualquer documento *web* acessível, analisa as opções de *crawling* disponíveis para encontrar *links* que, provavelmente, fornecerão conteúdo de alta relevância a um tópico definido *a priori*. Na abordagem de coleta de dados inteligente proposta neste trabalho, tópicos são definidos, não por palavras chaves, mas, pelo uso de documentos textuais como exemplos. Em seguida, técnicas de pré-processamento utilizadas em Mineração de Textos, entre elas o uso de um dicionário *thesaurus*, analisam semanticamente o documento apresentado como exemplo. Baseado nesta análise, o *web crawler* construído será guiado em busca do seu objetivo: recuperar informação relevante sobre o documento. A partir de sementes ou realizando uma consulta automática nas máquinas de buscas disponíveis, o *crawler* analisa, igualmente como na etapa anterior, todo documento recuperado na *Web*. Então, é executado um processo de comparação entre cada documento recuperado e o documento exemplo. Depois de obtido o nível de similaridade entre ambos, os *hyperlinks* do documento recuperado são analisados, empilhados e, futuramente, serão desempilhados de acordo seus

respectivos e prováveis níveis de importância. Ao final do processo de coleta de dados, outra técnica de Mineração de Textos é aplicada, objetivando selecionar os documentos mais representativos daquela coleção de textos: a Clusterização de Documentos. A implementação de uma ferramenta que contempla as heurísticas pesquisadas permitiu obter resultados práticos, tornando possível avaliar o desempenho das técnicas desenvolvidas e comparar os resultados obtidos com outras formas de recuperação de dados na *Web*. Com este trabalho, mostrou-se que o emprego de Mineração de Textos é um caminho a ser explorado no processo de recuperação de informação relevante na *Web*.

Palavras-chave

Mineração de Textos, Coleta de Dados, *Web Crawling*, Recuperação de Informação

Abstract

Soares, Fábio de Azevedo; Passos, Emmanuel Piseces Lopes (Advisor); Vellasco, Marley Maria Bernardes Rebuzzi (Advisor). **Text Mining at the intelligent web crawling process.** Rio de Janeiro, 2008. 120p. M.Sc. Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This dissertation presents a study about the application of Text Mining as part of the intelligent Web crawling process. The most usual way of gathering data in Web consists of the utilization of web crawlers. Web crawlers are softwares that, once provided with an initial set of URLs (seeds), start the methodical proceeding of visiting a site, store it in disk and extract its hyperlinks that will be used for the next visits. But seeking for content in this way is an expensive and exhausting task. An intelligent web crawling process, more than collecting and storing any web document available, analyses its available crawling possibilities for finding links that, probably, will provide high relevant content to a topic defined *a priori*. In the approach suggested in this work, topics are not defined by words, but rather by the employment of text documents as examples. Next, pre-processing techniques used in Text Mining, including the use of a Thesaurus, analyze semantically the document submitted as example. Based on this analysis, the web crawler thus constructed will be guided toward its objective: retrieve relevant information to the document. Starting from seeds or querying through available search engines, the crawler analyzes, exactly as in the previous step, every document retrieved in Web. the similarity level between them is obtained, the retrieved document's hyperlinks are analysed, queued and, later, will be dequeued according to each one's probable degree of importance. By the end of the gathering data process, another Text Mining technique is applied, with the propose of selecting the most representative document among the collected texts: Document Clustering. The implementation of a tool incorporating all the researched heuristics allowed to achieve results, making possible to evaluate the

performance of the developed techniques and compare all obtained results with others means of retrieving data in Web. The present work shows that the use of Text Mining is a track worthy to be exploited in the process of retrieving relevant information in Web.

Keywords

Text Mining, Data Retrieval, Web Crawling, Information Retrieval.

Sumário

1 Introdução	15
1.1. Motivação	15
1.2. Objetivos do Trabalho	19
1.3. Descrição do Trabalho	19
1.4. Organização da Dissertação	21
2 Mineração de Textos: Fundamentos	22
2.1. Definição	22
2.2. Principais Elementos	26
2.3. Documentos textuais são estruturados	28
2.4. Características representativas de um documento	30
2.5. Abordagens ao processo de Mineração de Textos	33
2.5.1. Análise Estatística	33
2.5.2. Análise Semântica	33
2.6. Áreas correlatas a Mineração de Textos	34
2.6.1. Ciência Cognitiva	35
2.6.2. Processamento de Linguagem Natural	36
2.6.3. Aprendizado de Máquina	36
2.6.4. Estatística	38
2.6.5. Recuperação de Informação	39
2.6.6. Mineração de Dados	40
3 Metodologia de Mineração de Textos	41
3.1. Coleta de Dados	42
3.2. Pré-Processamento	43
3.2.1. Tokenização	44
3.2.2. Remoção de <i>stopwords</i>	46
3.2.3. Processamento de Linguagem Natural	47
3.3. Indexação	51
3.3.1. Indexação Textual	52
3.3.2. Indexação Temática	53
3.4. Mineração	54
3.5. Análise	55
3.5.1. Precisão	57
3.5.2. Abrangência	57
3.5.3. Medida-F	57
3.5.4. Precisão x Abrangência	58
4 Recuperação de Informação	60
4.1. Introdução	60
4.2. Histórico da área de Recuperação de Informação	62
4.2.1. 1ª Fase – Décadas de 50 e 60	62
4.2.2. 2ª Fase – Décadas de 70 e 80	62

4.2.3. 3ª Fase – Década de 90 em diante	63
4.3. Recuperação de Informação Clássica	64
4.3.1. Modelo de Recuperação Booleano	66
4.3.2. Modelo de Espaço Vetorial	67
4.4. Recuperação de Informação na Internet	71
4.4.1. Crawlers	71
4.4.2. URL	74
4.4.3. Hiperlink	74
4.4.4. Políticas de <i>Web Crawling</i>	74
4.4.5. Máquinas de Busca	78
 5 <i>Crawler</i> Focado	 82
5.1. Definição	82
5.2. Exames sobre pesquisas na área	84
5.3. Estratégias	85
5.4. Web Analysis	86
5.5. Web Search	87
5.6. Estratégias Adicionais	87
 6 Metodologia Proposta	 89
6.1. Proposta	89
6.2. Metodologia Proposta	90
6.2.1. Módulo <i>Off-Line</i>	92
6.2.2. Módulo <i>On-Line</i>	98
6.3. Implementação	100
6.3.1. Ambiente de Desenvolvimento	100
6.3.2. Arquitetura Geral do Sistema	100
6.3.3. Multithreading	103
6.4. Estudo de Casos	103
 7 Conclusão e Trabalhos Futuros	 112
7.1. Conclusão	112
7.2. Trabalhos Futuros	113
 Referências Bibliográficas	 114

Lista de Figuras

Figura 1 – Número de servidores de internet registrados por ano.....	16
Figura 2 - Processo de obtenção de conhecimento.....	17
Figura 3 – Integridade semântica de um SGBD	24
Figura 4 – Sites brasileiros quanto à frequência de modificação do conteúdo.....	27
Figura 5 - Coleções de documentos com elementos em comum.....	28
Figura 6 – Algumas estruturas sintáticas de um trecho de texto	29
Figura 7 – Documentos em formatos fracamente estruturado e semi-estruturado (respectivamente).....	30
Figura 8 – Modelos de representação baseados em palavras e termos.....	32
Figura 9 – Multidisciplinaridade da Mineração de Textos.....	35
Figura 10 – Modelo simples de aprendizagem de máquina	37
Figura 11 – Teorema de Bayes	39
Figura 12 – Linhas cronológica das etapas de um processo de Mineração de Textos (por Aranha).....	41
Figura 13 – Processo de representação estruturada de um texto	44
Figura 14 - Metodologia de identificação de tokens proposta por KONCHADY	45
Figura 15 - Processo de tokenização seguido por remoção de stopwords.....	46
Figura 16 - Reconhecimento de anáfora com informações do contexto	48
Figura 17 – Erros de um processo de stemming: overstemming e understemming	50
Figura 18 – Derivações de um mesmo radical identificadas pelo algoritmo de Porter.....	51
Figura 19 - Representação de um índice invertido	53
Figura 20 - Estrutura básica de um Dicionário Thesaurus	54
Figura 21- Visualização das regras para concessão de empréstimo em uma árvore de decisão	56
Figura 22 – Fórmula da métrica de desempenho “Precisão”.....	57
Figura 23 – Fórmula da métrica de desempenho “Abrangência”.....	57
Figura 24 – Fórmula da métrica de desempenho “Medida-F”	58
Figura 25 – Gráfico de compensação entre precisão e abrangência.....	59
Figura 26 - Sistema Clássico de Recuperação de Informação	64
Figura 27 – Etapas possíveis no processo de Indexação de documentos textuais.....	66
Figura 28 - Representação Vetorial do Documento Di no espaço n-dimensional	68
Figura 29 – Cálculo da medida TF em um documento	69
Figura 30 – Cálculo da medida TF-IDF em um documento.....	70
Figura 31 – Similaridade entre dois documentos pela medida do Cosseno	71

Figura 32 – <i>Webgraph</i> do site do DEE/PUC-Rio.....	72
Figura 33 - Funcionamento de um crawler simples	73
Figura 34 - Ordem de visitas dos <i>sites</i> utilizando a estratégia <i>breadth-first</i>	77
Figura 35 – Pontuação do algoritmo de <i>pagerank</i>	77
Figura 36 - Processo de consulta em uma máquina de busca.....	79
Figura 37 – Arquitetura de uma máquina de busca moderna	80
Figura 38 – Arquitetura de um <i>crawler</i> focado dotado do elemento Destilador... 83	
Figura 39 - Relacionamentos de co-citação em uma comunidade <i>Web</i>	88
Figura 40 - Metodologia proposta para a Coleta Inteligente de Dados	92
Figura 41 - Módulo Off-Line da metodologia proposta.....	93
Figura 42 - Estrutura do dicionário Thesaurus utilizado no Sistema de MT.....	96
Figura 43 - Módulo On-Line da metodologia proposta.....	98
Figura 44 – Sequência de ações executadas na análise do documento exemplo	101
Figura 45 – Atividades desempenhadas pelo <i>crawler</i> para iniciar o rastreamento.....	101
Figura 46 - Procedimento iterativo de rastreamento do crawler	102
Figura 47 - Atividades executadas na etapa de Mineração de Textos.....	102
Figura 48 - Exemplo de documento do corpus CETENFolha.....	104
Figura 49 – Análise geral da distribuição de frequência de <i>tokens</i> de um documento	106
Figura 50 - Consultas realizadas com auxílio do dicionário Thesaurus	106
Figura 51 – Redução do número de tokens com <i>stemming</i>	107
Figura 52 - Documentos com erros de ortografia.....	108
Figura 53 - Relação entre similaridade e quantidade de termos utilizada na consulta realizada nos <i>search engines</i>	108
Figura 54 - Comparativo das técnicas de crawling	109
Figura 55 – Similaridade média dos documentos ao longo do processo de <i>crawling</i>	110
Figura 56 - Análise visual do processo de Clusterização	111

Lista de Tabelas

Tabela 1 - As duas abordagens para a Análise de Textos e suas principais Áreas de Conhecimento.....	34
Tabela 2 - As principais características de cada uma das abordagens para a Análise de Textos	34
Tabela 3 – Visualização das regras para concessão de empréstimos em uma tabela	56
Tabela 4 - Comparação entre Recuperação de Dados x Recuperação de Informação.....	61
Tabela 5 – Estatísticas de <i>crawling</i> do <i>site</i> do DEE.....	72
Tabela 6 - Lista de stopwords utilizadas na etapa de Pré-processamento	95
Tabela 7 - Informações adicionais sobre o CETENFolha_8 ____	104
Tabela 9 – Análise pontual da distribuição de frequências de <i>tokens</i> de um documento	105