Pontifícia Universidade Católica
do Rio de Janeiro

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

Pontifícia Universidade Católica
do Rio de Janeiro

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

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 Rebuzzi 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 Rebuzzi 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

Lista de Figuras

Lista de Tabelas