Pontifícia Universidade Católica
do Rio de Janeiro

Christian Nunes Aranha

Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional

Tese de Doutorado

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

Orientadora: Profa. Marley Maria Bernardes Rebuzzi Vellasco

Rio de Janeiro

Março de 2007

Pontifícia Universidade Católica
do Rio de Janeiro

Christian Nunes Aranha

Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional

Tese apresentada como requisito parcial para obtenção do grau de Doutor 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.

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

Dr. Emmanuel Piceses Lopes Passos
Co-Orientador
Departamento de Engenharia Elétrica - PUC-Rio

Dr. Marco Aurélio Cavalcanti Pacheco
Departamento de Engenharia Elétrica - PUC-Rio

Dr. Antonio Luz Furtado
Departamento de Informática – PUC-Rio

Dra. Maria Carmelita Padua Dias
Departamento de Letras – PUC-Rio

Dr. Nelson Francisco Favilla Ebecken
UFRJ

Dr. Alexandre Linhares
FGV-RJ

Dra. Valeria Menezes Bastos
Departamento de Informática – PUC-Rio

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

Rio de Janeiro, 28 de março de 2007

Christian Nunes Aranha

Graduado em Engenharia Elétrica, com ênfase em sistemas de apoio à decisão. Trabalho de final de curso em otimização lagrangeana aplicada ao problema de ordenação linear (LOP) sob a orientação de Abílio Lucena. Mestrado em métodos estatísticos de apoio à decisão. Dissertação sobre regressão multivariada linear por partes, modelo TS-TARX, sob a orientação de Alvaro Veiga. Doutorado em Inteligência Computacional para mineração de textos.

Ficha Catalográfica

Aranha, Christian Nunes

Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional / Christian Nunes Aranha; orientadora: Marley Maria Bernardes Rebuzzi Vellasco. – 2007.

144 f. ; 30 cm

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

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Mineração de texto. 3. Pré-processamento. 4. Inteligência artificial. I. Vellasco, Marley Maria Bernardes Rebuzzi. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

Agradecimentos

Ao professor Emmanuel Passos pelo seu apoio e orientação.

Aos meus pais, minha namorada e meus amigos, sem os quais são poderia ter completado este trabalho.

Ao CNPq pelo incentivo à pesquisa.

Resumo

Christian Nunes Aranha; Vellasco, Marley Maria Bernardes Rebuzzi (Orientadora). **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional**. Rio de Janeiro, 2007. 144p. Tese de Doutorado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

O presente trabalho apresenta uma pesquisa onde é proposto um novo modelo de pré-processamento para mineração de textos em português utilizando técnicas de inteligência computacional baseadas em conceitos existentes, como redes neurais, sistemas dinâmicos, e estatística multidimensional. O objetivo dessa tese de doutorado é, portanto, inovar na fase de pré-processamento da mineração de textos, propondo um modelo automático de enriquecimento de dados textuais. Essa abordagem se apresenta como uma extensão do tradicional modelo de conjunto de palavras (*bag-of-words)*, de preocupação mais estatística, e propõe um modelo do tipo conjunto de lexemas (*bag-of-lexems*) com maior aproveitamento do conteúdo lingüístico do texto em uma abordagem mais computacional, proporcionando resultados mais eficientes. O trabalho é complementado com o desenvolvimento e implementação de um sistema de pré-processamento de textos, que torna automática essa fase do processo de mineração de textos ora proposto. Apesar do objeto principal desta tese ser a etapa de pré-processamento, passaremos, de forma não muito aprofundada, por todas as etapas do processo de mineração de textos com o intuito de fornecer a teoria base completa para o entendimento do processo como um todo. Além de apresentar a teoria de cada etapa, individualmente, é executado um processamento completo (com coleta de dados, indexação, pré-processamento, mineração e pós-processamento) utilizando nas outras etapas modelos já consagrados na literatura que tiveram sua implementação realizada durante esse trabalho. Ao final são mostradas funcionalidades e algumas aplicações como: classificação de documentos, extração de informações e interface de linguagem natural (ILN).

Palavras-chave

Mineração de Texto, Preprocessamento, Inteligência Artificial.

Abstract

Christian Nunes Aranha; Vellasco, Marley Maria Bernardes Rebuzzi (Advisor). **An Automatic Preprocessing for Text Mining in Portuguese: A Computer-Aided Approach**. Rio de Janeiro, 2007. 144p. D.Sc. Thesis - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This work presents a research that proposes a new model of pre-processing for text mining in portuguese using computational intelligence techniques based on existing concepts, such as neural networks, dinamic systems and multidimensional statistics. The object of this doctoral thesis is, therefore, innovation in the pre-processing phase of text-mining, proposing an automatic model for the enrichment of textual data. This approach is presented as an extension of the traditional bag-of-words model, that has a more statistical emphasis, and proposes a bag-of-lexemes model with greater usage of the texts' linguistic content in a more computational approach, providing more efficient results. The work is complemented by the development and implementation of a text pre-processing system that automates this phase of th text mining process as proposed. Despite the object of this thesis being the pre-processing stage, one feels apropriate to describe, in overview, every step of the text mining process in order to provide the basic theory necessary to understand the process as a whole. Beyond presenting the theory of every stage individually, one executes a complete process (with data collection, indexing, pre-processing, mining and post-processing) using tried-and-true models in all the other stages, which were implemented during the development of this work. At last some functionalities and aplications are shown, such as: document classification, information extraction and natural language interface (NLI).

Keywords

Text Mining, Preprocessing, Artificial Inteligence.

Sumário

Lista de figuras

Lsta de tabelas