



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 Rebuszi Vellasco

Rio de Janeiro

Março de 2007



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 Rebuszi 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

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

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 não 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, Pré-processamento, 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 Intelligence.

Sumário

1. Introdução	13
1.1. Motivação	15
1.2. Objetivo da Tese	18
1.3. Estrutura da Tese	20
2. O Estado da Arte	22
2.1. Modelos Puramente Estatísticos	22
2.1.1. Modelo de Espaço Vetorial	23
2.1.2. Análise de Correspondência	25
2.1.3. Análise de Discriminante	28
2.2. Redes Neurais	31
2.2.1. Hopfield	31
2.2.2. Backpropagation	32
2.2.3. Mapas Auto-Organizáveis	33
2.3. Aprendizado de Máquina	35
2.3.1. Cadeias de Markov Escondidas	35
2.3.2. Aprendizado Baseado em Transformações	38
3. Mineração de Texto	40
3.1. Coleta de Dados	41
3.2. Pré-processamento	42
3.2.1. Identificação de Palavras no Texto	43
3.2.2. Redução de Dimensionalidade	45
3.2.3. Remoção de Palavras Não-Discriminantes (<i>Stop-words</i>)	46
3.3. Indexação	47
3.3.1. Procura Caractere à Caractere	49
3.3.2. Lista Invertida	50
3.3.3. Similaridade	52
3.3.4. Processo de Indexação	53
3.3.5. Índice do tipo Full-text	54
3.3.6. Ordenação	57
3.4. Mineração de Dados	58
3.5. Análise da Informação	59
4. Processamento da Linguagem Natural	60
4.1. O Modelo de PLN	60
4.1.1. Aquisição Automática	61
4.1.2. O Léxico	61
4.1.3. Sobre a Delimitação da Unidade Lexical	62
4.1.4. Ontologia	63
4.1.5. Precisão e Recordação	63
4.2. Técnicas de PLN	64
4.2.1. Tokenização	64
4.2.2. Normalização	65
4.2.3. Expressões Multi-Vocabulares	67

4.2.4. Fronteiras das Frases	68
4.2.5. Etiquetagem	69
4.2.6. Padrões Gramaticais	70
4.2.7. Reconhecimento de Entidades Mencionadas	70
4.2.8. Classificação de Entidades Mencionadas	73
4.2.9. Análise dos Constituintes	74
4.2.10. Correferência	74
4.2.10.1. Acrônimos, Siglas e Abreviaturas	75
4.2.10.2. Nomes Truncados	76
4.2.10.3. Anáfora Pronominal	77
4.2.10.4. Sinônimos	78
4.2.10.5. Erros Ortográficos	78
4.2.11. Discriminação do Sentido da Palavra	79
4.2.11.1. Detecção Automática de Sinônimos	80
5. Desenvolvimento e Implementação	82
5.1. Aprendizado Automático	82
5.2. O Léxico Computacional	83
5.2.1. A Importância do Léxico	84
5.3. Percepção Lingüística	85
5.4. Tesauro	87
5.5. O Modelo	88
5.5.1. Definições	89
5.5.2. Arquitetura	89
5.5.3. Compostos	99
5.5.4. Nomes Próprios	102
5.5.5. Sintaxe	102
5.6. Representação do Documento	103
5.6.1. Formato de Armazenamento	104
5.6.1.1. Tags de Categoria	104
5.6.1.2. Tags de Contexto	105
5.6.1.3. Tags de Função	106
5.6.1.4. Tags de Estrutura	107
5.6.1.5. Tags Descritivas	108
5.6.2. Exemplos	109
6. Exemplos de Aplicações de Mineração de Textos	111
6.1.1. Classificação	111
6.1.2. Extração de Informações	112
6.1.3. Interface em Linguagem Natural	114
6.2. Web Semântica	116
7. Resultados	119
7.1. Pré-processamento	119
7.2. Classificação	122
7.2.1. Amostra Pequena	123
7.2.2. Amostra Grande	127
8. Conclusão e Trabalhos Futuros	129

9. Referências bibliográficas	131
10. Anexo I: Principais distribuições de frequência dos significados	142
11. Anexo II: Exemplo de Mineração de Textos por Perguntas	143

Lista de figuras

Figura 1 – Valor agregado pela Mineração de Textos (MT) no processo de análise de informações textuais	17
Figura 2 – Diagrama de camadas abstratas de um sistema computacional	19
Figura 3 – Resultado em duas dimensões da análise de correspondência	27
Figura 4 – Análise de correspondência utilizando a etapa de pré com segmentos de frase	28
Figura 5 – Arquitetura de Hopfield utilizada no software TextAnalyst	32
Figura 6 – Rede Neural Backpropagation. Modelo usado por Fukuda	33
Figura 7 – Distribuição do mapa dos documentos. Em cinza os lugares onde há a maior concentração de documentos	34
Figura 8 – Exemplo de automato para Cadeia de Markov Escondida	36
Figura 9 – Texto original	37
Figura 10 – Resultado da extração de informação de HMM. Os rótulos após cada palavra é o rótulo ótimo encontrado. <T> título; <D> data; <A> autor e <C> corpo	37
Figura 11 – Processo de funcionamento de um TBL	38
Figura 12 – Identificação de palavras válidas	44
Figura 13 – A curva de Zipf e os cortes de Luhn	46
Figura 14 – Identificação de <i>Stop-Words</i>	47
Figura 15 – Estrutura de uma Lista Invertida associada aos documentos indexados.	51
Figura 16 - Função Similaridade	52
Figura 17 – Sequência do processo de indexação automática.	54
Figura 18 – Arquitetura da Busca tipo Full-text	56
Figura 19 – Representação de uma estrutura de hiperlinks na Internet	57
Figura 20 – Os três gráficos (1), (2) e (3) mostram, de forma ilustrativa, a necessidade colaborativa de três tarefas de PLN T1, T2 e T3. T2 só consegue atingir 90% de acerto de melhorar T1 e T3.	60
Figura 21 – As figures (a), (b) e (c) ilustram três relações diferentes e hipotéticas de sinonímia.	67
Figura 22 – Esquema de pré-requisitos entre as classes ontológicas.	73
Figura 23 – Esquema geral de um extrator de informações lingüísticas	83
Figura 24 – Modelo de banco de dados.	90
Figura 25 – Modelo de classes por orientação a objeto.	90
Figura 26 – (<i>Pipeline</i>) Sequência de procedimentos de reconhecimento de padrões e aprendizado de lexemas especializados em cada área do PLN.	91
Figura 27 – Modelo de aprendizado autonomo e retroalimentado.	92
Figura 28 – Sequências de etapas que compõe o processamento do texto	93
Figura 29 – Ontologia pré-definida para o processamento. As setas indicam os pré-requisitos ($A \rightarrow B = A$ é pré-requisito de B)	93
Figura 30 – Cada círculo representa um lexema percebido. A borda de cada um deles indica um traço semântico. O resultado das referências agrupa as redundâncias semânticas	95

Figura 31 – Exemplo de uma gramática escrita sobre a de especificação de Backus-Norm-Form	97
Figura 32– Suporte da interface sintática forçando a reavaliação da percepção e referência	98
Figura 33 – Diferentes módulos para resolver a sintaxe. Estão separados por fatores de performance e forma de execução	103
Figura 34 – Exemplos de tags de categoria	105
Figura 35 – Exemplo de Tags de Contexto	106
Figura 36 – Exemplos de Tags de Função	107
Figura 37 – Exemplos de Tags Descritivas	108
Figura 38 – Exemplo de reconhecimento de entidades. Instituições em vermelho, índice em verde e tempo em roxo	109
Figura 39 – Exemplo de reconhecimento de entidades. Nomes de pessoas em laranja e lugares em azul.	109

Lsta de tabelas

Tabela 1 – Ranking de quantidade de Hosts por país	17
Tabela 2 – Matriz com o número de ocorrência do termo (linhas) em cada uma das partições educação (L,M e H) e idade (-30, -55 e +55).	26
Tabela 3 – Especificação da tabela e exemplos de registros do léxico utilizado	62
Tabela 4 – Resultado do Reconhecimento de Entidades Mencionadas para os textos da coleção escritos no Português Brasileiro	120
Tabela 5 – Resultado HAREM para a avaliação de bases de emails	121
Tabela 6 – Resultado HAREM para a avaliação de bases de textos de jornais	122
Tabela 7 – Resultado HAREM para a avaliação de bases de páginas da Internet	122
Tabela 8 – Resultados do modelo bag-of-words para a fase de treinamento com 5000 notícias. A tabela apresenta os dados do modelo e o quadro apresenta o formato de leitura dos parâmetros do modelo	124
Tabela 9 – Resultados do modelo bag-of-lexems para a fase de treinamento com 5000 notícias. A tabela apresenta os dados do modelo e o quadro apresenta o formato de leitura dos parâmetros do modelo	124
Tabela 10 – Resultados do modelo bag-of-lexems com ontologia para a fase de treinamento com 5000 notícias. A tabela apresenta os dados do modelo e o quadro apresenta o formato de leitura dos parâmetros do modelo	125
Tabela 11 – Comparação entre as acurácias dos tres métodos: BOW (bag-of-words), BOL(bag-of-lexems) e BOLO(bag-of-lexems com ontologia), para as diferentes quantidades de informação processada	126
Tabela 12 – Comparativo do número de termos significantes usados no modelo de classificação	126
Tabela 13 – Resultados de performance e generalização utilizando bag-of-words	127
Tabela 14 – Resultados de performance e generalização utilizando bag-of-lexems	128
Tabela 15 – Resultados de performance e generalização utilizando bag-of-lexems rotulados usando a ontologia definida	128