Pontifícia Universidade Católica
do Rio de Janeiro

Roberto Miranda Gomes

Desambiguação de Sentido de Palavras Dirigida por Técnicas de Agrupamento sob o Enfoque da 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.

Orientadora: Profa. Marley Maria Bernades Rebuzzi Vellasco

Rio de Janeiro

Março de 2009

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Roberto Miranda Gomes

Desambiguação de Sentido de Palavras Dirigida por Técnicas de Agrupamento sob o Enfoque da Mineração de Textos

Dissertação apresentada como requisito parcial para obtenção do grau de 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 Bernades Rebuzzi Vellasco
Orientadora
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Emmanuel Piceses Lopes Passos
Co-orientador
PUC-Rio

Prof. Antonio Luz Furtado
Departamento de Engenharia de Informática - PUC-Rio

Prof. Christian Nunes Aranha
PUC-Rio

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico

Rio de Janeiro, 05 de março de 2009

Roberto Miranda Gomes

Graduou-se em Engenharia de Computação no Instituto Militar de Engenharia (IME). Cursou o Mestrado em Sistemas e Computação do IME onde ministrou aulas e contribuiu com a orientação de trabalhos na área de Segurança da Informação para alunos de graduação. Trabalhou em diversos projetos relacionados à Segurança da Informação dentro do Exército Brasileiro e Governo Federal (incluindo a ICP-Brasil) e hoje possui uma empresa que presta consultoria de inteligência nos negócios para apoio à decisões estratégicas

Ficha Catalográfica

Gomes, Roberto Miranda

Desambiguação de sentido de palavras dirigida por técnicas de agrupamento sob o enfoque da mineração de textos / Roberto Miranda Gomes ; orientadora: Marley Maria Bernades Rebuzzi Vellasco; co-orientador: Emmanuel Piceses Lopes Passos. – 2009.

118 f. ; 30 cm

Dissertação (Mestrado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2009.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Mineração de textos. 3. Desambiguação. 4. Técnicas de agrupamento. I. Vellasco, Marley Maria Bernades Rebuzzi. II. Passos, Emmanuel Piceses Lopes. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Dedicado à minha família.

Agradecimentos

À PUC pelo excelente ambiente e pelos recursos disponibilizados e a CAPES pelo financiamento dessa pesquisa.

Aos meus pais e irmãs que, com muito carinho e apoio, não mediram esforços para que eu concluísse mais essa etapa dos meus estudos.

Ao professores e orientadores Marley Vellasco e Emmanuel Passos pele apoio e inspiração no amadurecimento dos meus conhecimentos e conceitos que me levaram a execução e conclusão deste trabalho. Particularmente ao professor Emmanuel ainda, pela paciência na orientação e incentivo constante.

Aos colegas, Fábio Azevedo e João Carrilho, pelo convívio, pelo apoio, pela compreensão e pela amizade.

A todos os demais professores da PUC que ministraram aulas para mim e muito contribuíram com seus conhecimentos e estímulo, em mim despertado, pela maestria com a qual conduziram os ensinamentos. Em especial ao professor Antonio Furtado pelo exemplo de humildade e demonstração de interesse pelo trabalho.

Às secretárias do Departamento de Engenharia Elétrica, Alcina e Marcia, pelo convívio e prestatividade.

Resumo

Gomes, Roberto Miranda; Vellasco, Marley Maria Bernades Rebuzzi (Orientadora); Passos, Emmanuel Piceses Lopes (Co-orientador). **Desambiguação de Sentido de Palavra Dirigida por Técnicas de Agrupamento sob o Enfoque da Mineração de Textos**. Rio de Janeiro, 2009, 118p. Dissertação de mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Esta dissertação investigou a aplicação de processos de mineração de textos a partir de técnicas de inteligência computacional e aprendizado de máquina no problema de ambigüidade de sentido de palavras. O trabalho na área de métodos de apoio à decisão teve como objetivo o desenvolvimento de técnicas capazes de automatizar os processos de desambiguação bem como a construção de um protótipo baseado na implementação de algumas dessas técnicas. Desambiguação de sentido de palavra é o processo de atribuição de um significado a uma palavra obtido por meio de informações colhidas no contexto em que ela ocorre, e um de seus objetivos é mitigar os enganos introduzidos por construções textuais ambíguas, auxiliando assim o processo de tomada de decisão. Buscou-se ainda na utilização de conceitos, ferramentas e formas de documentação considerados em trabalhos anteriores de maneira a dar continuidade ao desenvolvimento científico e deixar um legado mais facilmente reutilizável em trabalhos futuros. Atenção especial foi dada ao processo de detecção de ambigüidades e, por esse motivo, uma abordagem diferenciada foi empregada. Diferente da forma mais comum de desambiguação, onde uma máquina é treinada para desambiguar determinado termo, buscou-se no presente trabalho a não-dependência de se conhecer o termo a ser tratado e assim tornar o sistema mais robusto e genérico. Para isso, foram desenvolvidas heurísticas específicas baseadas em técnicas de inteligência computacional. Os critérios semânticos para identificação de termos ambíguos foram extraídos das técnicas de agrupamento empregadas em léxicos construídos após algum processo de normalização de termos. O protótipo, SID – Sistema Inteligente de Desambiguação – foi desenvolvido em .NET, que permite uma grande diversidade de linguagens no desenvolvimento, o que facilita o reuso do código para a continuidade da pesquisa ou a utilização das técnicas implementadas em alguma aplicação de mineração de textos. A linguagem escolhida foi o C#, pela sua robustez, facilidade e semelhança sintática com JAVA e C++, linguagens amplamente conhecidas e utilizadas pela maioria dos desenvolvedores.

Palavras-chave

Mineração de textos, Desambiguação, Técnicas Agrupamento.

Abstract

Gomes, Roberto Miranda; Vellasco, Marley Maria Bernades Rebuzzi (Advisor); Passos, Emmanuel Piceses Lopes (Co-advisor). **Word Sense Desambiguation in Text Mining**. Rio de Janeiro, 2009, 118p. MSc Dissertation – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This dissertation investigated the application of text mining process from techniques of computing intelligence and machine learning in the problem of word sense ambiguity. The work in the methods of decision support area aimed to develop techniques capable of doing a word meaning disambiguation automatically and also to construct a prototype based on the application of such techniques. Special attention was given to the process of ambiguity detection and, for this reason, a differentiated approach was used. Unlikely the most common type of disambiguation, in which the machine is trained to do it in determined terms, the present work aimed to address the ambiguity problem without the need of knowing the meaning of the term used, and thus, to make the system more robust and generic. In order to achieve that, specific heurists were developed based on computing intelligence techniques. The semantic criteria used to identify the ambiguous terms were extracted from grouping techniques employed in lexis built after some term normalization process.

Keywords

Desambiguation, Text Mining, Clustering.

"A vida é como jogar uma bola na parede: se for jogada uma bola azul, ela voltará azul; se for jogada uma bola verde, ela voltará verde; se for jogada fraca, ela voltará fraca; se for jogada com força, ela voltará com força. Por isso, nunca jogue uma bola na vida de forma que você não esteja pronto a recebê-la. A vida não dá nem empresta; não se comove nem se apieda. Tudo quanto ela faz é retribuir e transferir aquilo que nós lhe oferecemos."

Albert Einstein

Sumário

Introdução 13

1. Mineração de textos 16
   1.1 O que a Mineração de Textos é capaz de fazer 17
   1.2 Elementos básicos e estruturação 18
   1.3 Ferramentas para Mineração de Textos 23
      1.3.1 Classificação de Documentos 23
      1.3.2 Agrupamento de Documentos 25
      1.3.3 Predição e Avaliação 27
      1.3.4 Recuperação de Informações 28
      1.3.5 Extração de Informações 30

2. Processo de Mineração de Textos 32
   2.1 Coleta 33
   2.2 Pré-processamento 34
      2.2.1 Tokenização 36
      2.2.2 Processamento de Linguagem Natural 37
         2.2.2.1 Identificação de colocações 38
         2.2.2.2 Classes gramaticais 38
         2.2.2.3 Lematização 39
         2.2.2.4 Análise de discurso 41
   2.3 Indexação 42
   2.4 Mineração 44
   2.5 Análise 45

3. Recuperação de Informações Textuais 46
   3.1 Modelo Booleano 49
   3.2 Modelo Probabilístico 50
   3.3 Modelo vetorial 51
   3.4 Extensões aos modelos 54

4. Agrupamento de documentos 58
   4.1 Escolha das variáveis 59
   4.2 Método de formação dos grupos 60
   4.3 Métricas para agrupamento 62
      4.3.1 Métricas de Distância 63
      4.3.2 Métricas de Similaridade 64

5. Ambigüidades no contexto de Mineração de Textos 66
   5.1 Tipos de ambigüidades 66
   5.2 Desambiguação de sentido de palavra 68
   5.3 Pesquisas na área de desambiguação 69
   5.4 Modelo geral para tratamento de ambigüidades 73

6. Sistema Proposto 76
   6.1 Desambiguação baseada em agrupamento de documentos 76
   6.2 Processamento de textos 79

6.2.1 Indexação do Corpus 81
6.2.2 Indexação do Léxico 81
6.2.3 Montagem do modelo vetorial 82

7. Estudo de Caso 84
7.1 Montagem do Corpus 84
7.2 Resultados 85
7.3 Trabalhos futuros 102
7.4 Conclusões 104

Referências bibliográficas 107

Lista de abreviaturas

FTC	*Frequent Term-based Clustering*
HFTC	*Hierarchical Frequent Term-based Clustering*
HIFC	*Frequent Itemset-based Hierarchical Clustering*
HMM	*Hidden Markov Model*
HTML	*HiperText Markup Language*
KDD	*Knowledge Discovery in Data*
KDT	*Knowledge Discovery in Text*
MUC	*Message Understanding Conference*
NIST	*National Institute of Standards and Technology*
PLN	Processamento de Linguagem Natural
POS	*Part of Speech*
RNA	Redes Neurais Artificiais
SGBD	Sistema de Gerenciamento de Banco de Dados
SSI	*Structural Semantic Interconnections*
STC	*Suffix Tree Clustering*
SVM	*Support Vector Machine*
TBL	*Tranformation Based Learning*
WBSC	*Word-based Soft Clustering*
XML	*eXtended Markup Language*
WSD	*Word Sense Desambiguation*

Lista de figuras

Figura 1 – Organização hierárquica com o tipo em nível superior ao assunto. 19
Figura 2 – Organização hierárquica com o assunto em nível superior ao tipo. 19
Figura 3 – Tabelas da base de dados Loja. 21
Figura 4 – Documento XML representando a base de dados Loja. 22
Figura 5 – Classificação ternária de documentos. 24
Figura 6 – Agrupamento não-hierárquico de documentos em 3 grupos. 25
Figura 7 – Agrupamento de não-hierárquico aglomerativo de documentos (dendrograma). 26
Figura 8 – Processo de predição pelo aprendizado de máquina. 28
Figura 9 – Processo de recuperação da informação. 29
Figura 10 – Extração de informações de um documento. 30
Figura 11 – Etapas do processo de extração de conhecimento em textos. 32
Figura 12 – Etapas gerais do processo de pré-processamento. 36
Figura 13 – Taxonomia para métodos de redução ao radical 40
Figura 14 – Processo de mineração de dados em textos. 44
Figura 15 – Cosseno de θ, adotado como medida de similaridade. 52
Figura 16 – Matriz representativa do corpus constituído pelos três documentos apresentados 58
Figura 17 – Distância baseada na sobreposição de áreas. 64
Figura 18 – Modelo geral para tratamento de ambigüidades 74
Figura 19 – Indexação de documentos com informações de agrupamento. 77
Figura 20 – Recuperação de informações desambiguadas por informações de agrupamento 78
Tabela 1 – Codificação das classes naturais do corpus Cetenfolha 80
Figura 21 – Índice invertido, onde termos são índices para documentos. 82
Figura 22 – Extrato do léxico 1 indexado 87
Figura 23 – Dendrograma nº.1 de clusterização hierárquica aglomerativa 88
Figura 24 – Extrato do léxico 2 indexado 89
Figura 25 – Dendrograma nº.2 de clusterização hierárquica aglomerativa 91
Figura 26 – Extrato do léxico 3 indexado 92
Figura 27 – Dendrograma nº.3 de clusterização hierárquica aglomerativa 93
Tabela 2 – Resumo dos resultados do agrupamento hierárquico 94
Figura 28 – Gráfico silhouette do experimento 1 com 2 grupos 95
Figura 29 – Gráfico de soma das distâncias para o experimento 1 96
Figura 30 – Gráfico silhouette do experimento 1 com 11 grupos 97
Figura 31 – Gráfico silhouette do experimento 2 com 2 grupos 98
Figura 32 – Gráfico de soma das distâncias para o experimento 2 99
Figura 33 – Gráfico silhouette do experimento 2 com 12 grupos 99
Figura 34 – Gráfico silhouette do experimento 3 com 2 grupos 100
Figura 35 – Gráfico de soma das distâncias para o experimento 3 101
Figura 36 – Gráfico silhouette do experimento 3 com 7 grupos 101
Figura 37 – Função de normalização, que mapeia palavras em termos 105