

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO



Susana Rosich Soares Velloso

**SQLLOMining: Obtenção de Objetos de Aprendizagem
utilizando técnicas de Aprendizado de Máquina**

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para
obtenção do título de Mestre pelo Programa de Pós-
Graduação em Informática da PUC-Rio.

Orientador: Rubens Nascimento Melo

Rio de Janeiro, Julho de 2007



Susana Rosich Soares Velloso

SQLLOMining: Obtenção de Objetos de Aprendizagem utilizando técnicas de Aprendizado de Máquina

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Rubens Nascimento Melo
Orientador
PUC-Rio

Ruy Luiz Milidiú
Departamento de Informática - PUC-Rio

Sérgio Lifschitz
Departamento de Informática - PUC-Rio

José Eugênio Leal
Coordenador(a) Setorial do Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 20 de julho de 2007

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

Susana Rosich Soares Velloso

Graduou-se em Engenharia Mecânica na PUC-Rio em 1994. Atuou como consultora em Banco de Dados para diversas empresas. Possui interesse acadêmico e profissional nas áreas de Inteligência Artificial e Banco de Dados.

Ficha Catalográfica

Velloso, Susana Rosich Soares

SQLLOMining: Obtenção de objetos de aprendizagem utilizando técnicas de aprendizado de máquina / Susana Rosich Soares Velloso; orientador: Rubens Nascimento Melo. – 2007.

118 f.; 30 cm

Dissertação (Mestrado em Informática)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

Inclui bibliografia

1. Informática – Teses. 2. Educação baseada na Web. 3. Objetos de aprendizagem. 4. Ontologia. 5. Banco de dados para e-Learning. 6. Aprendizado de máquina. 7. Classificação de textos. 8. Naive-Bayes. I. Melo, Rubens Nascimento. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Este trabalho é dedicado ao meu marido
pelo amor e companheirismo.

Agradecimentos

À PUC-Rio, ao departamento de informática e ao CAPES pela oportunidade.

Ao meu orientador, Rubens Nascimento Melo, pela motivação, paciência e ajuda.

Aos meus companheiros e companheiras de TecBD, pela motivação e ajuda.

A todos os professores, funcionários do Departamento de Informática pelo apoio dado quando precisei.

Ao professor Ruy Luiz Milidiú pelo apoio e pelo curso "Text Mining", grande inspiração para a minha dissertação.

Ao professor Eduardo Sany Laber pelo curso de PAA e pela carta de recomendação para o meu ingresso no departamento.

Às minhas amigas pelos momentos de pura diversão e pelo incentivo nos momentos difíceis.

Aos meus familiares pela educação e pelo apoio.

Aos professores que participaram da banca examinadora.

Resumo

Velloso, Susana Rosich Soares. **SQLLOMining: Obtenção de Objetos de Aprendizagem utilizando técnicas de Aprendizado de Máquina**. Rio de Janeiro, 2007. 118p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Objetos de Aprendizagem ou Learning Objects (LOs) são porções de material didático tais como textos que podem ser reutilizados na composição de outros objetos maiores (aulas ou cursos). Um dos problemas da reutilização de LOs é descobri-los em seus contextos ou documentos texto originais tais como livros, e artigos. Visando a obtenção de LOs, este trabalho apresenta um processo que parte da extração, tratamento e carga de uma base de dados textual e em seguida, baseando-se em técnicas de aprendizado de máquina, uma combinação de EM (Expectation-Maximization) e um classificador Bayesiano, classifica-se os textos extraídos. Tal processo foi implementado em um sistema chamado "SQLLOMining", que usa SQL como linguagem de programação e técnicas de mineração de texto na busca de LOs.

Palavras-chave

Educação Baseada na Web; Objetos de Aprendizagem; Ontologia; Banco de Dados para e-Learning; Aprendizado de Máquina; Classificação de Textos; Naive-Bayes

Abstract

Velloso, Susana Rosich Soares. **SQLLOMining: Finding Learning Objects using machine learning methods**. Rio de Janeiro, 2007. 118p. MSc. Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Learning Objects (LOs) are pieces of instructional material like traditional texts that can be reused in the composition of more complex objects like classes or courses. There are some difficulties in the process of LO reutilization. One of them is to find pieces of documents that can be used like LOs. In this work we present a process that, in search for LOs, starts by extracting, transforming and loading a text database and then continue clustering these texts, using a machine learning methods that combines EM (Expectation-Maximization) and a Bayesian classifier. We implemented that process in a system called "SQLLOMining" that uses the SQL language and text mining methods in the search for LOs.

Keywords

Web Based Education; Learning Objects; Ontology; Database; e-Learning; Machine Learning; Text Classification; Naive-Bayes; Internet.

Sumário

1. Introdução	14
1.1. Motivação	16
1.2. Objetivo	16
1.3. Organização	17
2. Objetos de Aprendizagem	18
2.1. LOs / ALOs	18
2.2. Definição das classes	21
2.3. Ontologia de tipos de ALOs	23
3. Aprendizado de Máquina	29
3.1. Tipos de Aprendizado	30
3.2. Algoritmo EM Bayesiano de Misturas Multinomiais	31
3.2.1. Modelo Bayesiano	31
3.2.2. Classificador naive-Bayes	32
3.2.3. Distribuição Multinomial	33
3.2.4. Modelo Bayesiano de Misturas Multinomiais	34
3.2.5. O Algoritmo EM	35
3.2.6. Pseudocódigo	36
3.2.7. Suavização de Laplace	38
3.2.8. Métricas <i>precision-recall</i> e F1	38
4. SQLLOMining - um sistema de mineração de LOs	41
4.1. Especificação do processo de extração de Objetos de Aprendizado	44
4.1.1. Geração do Corpus Inicial	44
4.1.1.1. Definição dos tipos de ALOs	45
4.1.1.2. Carga de arquivos	45
4.1.1.3. Fragmentação do texto	47
4.1.2. Geração do Modelo Inicial	47

4.1.2.1. Pré-processamento do Corpus	48
4.1.2.2. Algoritmo de aprendizado de máquina	51
4.1.2.3. Aprendizado Semi-Supervisionado	52
4.1.3. Carga do Arquivo a ser Classificado	53
4.1.3.1. Carga e Fragmentação	53
4.1.3.2. Aprendizado Semi-supervisionado	55
4.1.4. Pesquisa de ALOs	55
4.2. Diagramas de Especificação	57
4.2.1. Diagrama de Classes	58
4.2.2. Diagramas de Modelagem	59
5 . Estudo de Casos	62
5.1. Geração do Corpus	62
5.1.1. Extração automática de exemplos	63
5.1.2. Pré-etiquetagem com conferência manual	64
5.2. Descrição dos experimentos	64
5.2.1. Corpus Definição	64
5.2.2. Corpus Estendido	65
5.3. Resultados	67
5.3.1. Corpus Definição	67
5.3.1.1. Seleção de <i>features</i> do modelo	67
5.3.1.2. Aprendizado Semi-Supervisionado	68
5.3.1.3. Classificação de arquivos	69
5.3.2. Corpus Estendido	70
5.3.2.1. Desempenho da Classificação	71
5.3.2.2. Desempenho do Aprendizado Semi-Supervisionado	73
6 . Trabalhos Relacionados	75
6.1. Descrição dos trabalhos	76
6.2. Comparação com outros trabalhos	79
7 . Conclusão	83
7.1. Contribuições	84
7.2. Trabalhos Futuros	84
Referências bibliográficas	86

Apêndice I	89
Funções para Algoritmo Porter	89
Tabela Stemming	94
Apêndice II	98
Procedures SQL para Algoritmo EM	98

Lista de figuras

Figura 1- Modelo reprodutor de LOs (Pereira, Porto e Melo, 2003)	20
Figura 2 - RIO	21
Figura 3 - RLO	22
Figura 4 - Ontologia de Objetos de Aprendizagem (Ullrich Carsten 2004, 2005)	25
Figura 5 - Geração do Corpus Inicial	42
Figura 6 - Geração do Modelo Inicial	42
Figura 7 - Carga do Arquivo a ser classificado	43
Figura 8 - Aprendizado Semi-Supervisionado	43
Figura 9 - Pesquisa de ALOs	44
Figura 10 - Tela de Definição de Tipos de ALOs	45
Figura 11 - Tela Carga de Arquivos	46
Figura 12 – Tela Pré-Processamento do Corpus	48
Figura 13 – Tela Geração do Modelo Multinomial	52
Figura 14 – Tela Aprendizado Semi_supervisionado	53
Figura 15 – Tela Experimento por Arquivo - Carga do Arquivo	53
Figura 16 – Tela Experimento por Arquivo – Fragmentação do Texto	54
Figura 17 – Tela Experimento por Arquivo – Pré-rocessamento	54
Figura 18 - Tela Experimento por Arquivo – Aprendizado Semi-Supervisionado	55
Figura 19 – Tela Pesquisa de ALOs	56
Figura 20 – Tela Pesquisa de Sentenças	57
Figura 21 - Diagrama de Classes	58
Figura 22- Diagrama ER	59
Figura 23- Modelo Lógico	61
Figura 24 – Accuracy Corpus Definição	68
Figura 25- Accuracy Corpus Aumentado	69
Figura 26 - Comparação da curva de Aprendizado - Valores de Accuracy	73
Figura 27 – <i>Precision</i> e <i>Recall</i> para Classe Definição	74
Figura 28– <i>Precision</i> e <i>Recall</i> para Classe Lei	74
Figura 29 – <i>Precision</i> e <i>Recall</i> para Classe Negativa	74

Lista de tabelas

Tabela 1 – Saco de Palavras Unigrama	49
Tabela 2 – Saco de Palavras Bigrama	50
Tabela 3 – Resultados com o Corpus inicial avaliação de <i>features</i>	67
Tabela 4 – Ganhos de desempenho com o Aprendizado Semi-Supervisionado	69
Tabela 5 – Resultados com o Corpus Definição	70
Tabela 6 – Comparação do Corpus Definição com o Corpus Estendido	71
Tabela 7 – Distribuição dos exemplos nos Corpus	72
Tabela 8 – Comparação do Corpus Definição com o Corpus Estendido	72
Tabela 9 – Resultados do primeiro trabalho relacionado	77
Tabela 10 – Resultados do segundo trabalho relacionado	78
Tabela 11 - Resultados do terceiro trabalho relacionado	79
Tabela 12 – Comparação de resultados com segundo trabalho relacionado	81
Tabela 13 – Resultados obtidos com o Corpus Lei variando a quantidade de exemplos	82

O conhecimento é a pequena porção da ignorância que arrumamos e
classificamos.

Ambrose Bierce