Pontifícia Universidade Católica
do Rio de Janeiro

David Steinbruch

Um estudo de algoritmos para classificação automática de textos utilizando naive-Bayes

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: Daniel Schwabe
Co-orientador: Ruy Luiz Milidiú

Rio de Janeiro, setembro de 2006

Pontifícia Universidade Católica
do Rio de Janeiro

David Steinbruch

Um estudo de algoritmos para classificação automática de textos utilizando naive-Bayes

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.

Daniel Schwabe
Orientador
PUC-Rio

Ruy Luiz Milidiú
Co-orientador
PUC-Rio

Marcus Vinicius Soledade Poggi de Aragão
PUC-Rio

Eduardo Sany Laber
PUC-Rio

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

Rio de Janeiro

05 de setembro de 2006

David Steinbruch

Graduou-se em Engenharia de Computação na PUC-Rio em 2003. Atuou como programador no desenvolvimento de soluções web e aplicações de mineração de textos. Possui interesse acadêmico e profissional nas áreas de Inteligência Artificial, Hipertexto e Multimídia.

Ficha Catalográfica

Steinbruch, David

Um estudo de algoritmos para classificação automática de textos utilizando naive-Bayes / David Steinbruch ; orientador: Daniel Schwabe. – 2006.

78 f. : il. ; 30 cm

Dissertação (Mestrado em Informática)– Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2006.

Inclui bibliografia

1. Informática – Teses. 2. Aprendizado de máquina. 3. Categorização de textos. 4. Classificação de textos. 5. Multirótulo. 6. Naive-Bayes. 7. Internet. I. Schwabe, Daniel. II. Milidiú, Ruy Luiz. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

Este trabalho é dedicado a minha família,
pela compreensão e pelo apoio,
e ao meu orientador e coorientador,
pela motivação.

Agradecimentos

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

Ao meu orientador, Daniel Schwabe, e co-orientador, Ruy Luiz Milidiú, pela motivação, paciência, confiança e ajuda.

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

A Deborah e Emanuelle pela paciência e ajuda nos processos burocráticos.

Aos meus amigos que cursaram o mestrado comigo e familiares pelo incentivo e apoio nos momentos difíceis.

Aos meus irmãos Daniel, Rachel e Natan pelos ótimos momentos juntos.

Aos meus pais Luna e Beni e meu padrasto Samy por tudo.

Resumo

Steinbruch, David; Schwabe, Daniel. **Um estudo de algoritmos para classificação automática de textos utilizando naive-Bayes**. Rio de Janeiro, 2006. 78p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A quantidade de informação eletrônica vem crescendo de forma acelerada, motivada principalmente pela facilidade de publicação e divulgação que a Internet proporciona. Desta forma, é necessária a organização da informação de forma a facilitar a sua aquisição. Muitos trabalhos propuseram resolver este problema através da classificação automática de textos associando a eles vários rótulos (classificação multirótulo). No entanto, estes trabalhos transformam este problema em subproblemas de classificação binária, considerando que existe independência entre as categorias. Além disso, utilizam limiares ("thresholds"), que são muito específicos para o conjunto de treinamento utilizado, não possuindo grande capacidade de generalização na aprendizagem. Esta dissertação propõe dois algoritmos de classificação automática de textos baseados no algoritmo multinomial naive Bayes e sua utilização em um ambiente on-line de classificação automática de textos com realimentação de relevância pelo usuário. Para testar a eficiência dos algoritmos propostos, foram realizados experimentos na base de notícias Reuters 21758 e na base de documentos médicos Ohsumed.

Palavras-chave

Aprendizado de Máquina; Categorização de Textos, Classificação de Textos; Multirótulo; Naive-Bayes; Internet

Abstract

Steinbruch, David; Schwabe, Daniel**. A study of multilabel text clasification algorithms using naive-Bayes.** Rio de Janeiro, 2006. 78p. MSc. Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The amount of electronic information has been growing fast, mainly due to the easiness of publication and spreading that Internet provides. Therefore, is necessary the organisation of information to facilitate its retrieval. Many works have solved this problem through the automatic text classification, associating to them several labels (multilabel classification). However, those works have transformed this problem into binary classification subproblems**,** considering there is not dependence among categories. Moreover, they have used thresholds, which are very sepecific of the classifier document base, and so, does not have great generalization capacity in the learning process. This thesis proposes two text classifiers based on the multinomial algorithm naive Bayes and its usage in an on-line text classification environment with user relevance feedback. In order to test the proposed algorithms efficiency, experiments have been performed on the Reuters 21578 news base, and on the Ohsumed medical document base.

Keywords

Machine Learning; Text Categorization; Text Classification; Multilabel; Naive-Bayes; Internet

Sumário

Siglas

Micro Recall	Micro Averaged Recall
Macro Recall	Macro Averaged Recall
Micro Precision	Micro Averaged Precision
Macro Precision	Macro Averaged Precision
Micro F1	Micro Averaged F1
Macro F1	Macro Averaged F1
MESH	Medical Subject Heading
R(10)	Subconjunto de 10 categorias da base de dados Reuters 21758
R(90)	Subconjunto de 90 categorias da base de dados Reuters 21758
CONSTRUE	Categorization of News STories, Rapidly, Uniformly and Extensibly
TF	Term frequency
IDF	Inverse Document Frequencie
Reuters 21578	Base de notícias da Reuters com 21578 documentos
Ohsumed	Subconjunto de 348.566 documentos da base MEDLINE
MEDLINE	Medical Literature Analysis and Retrieval System Online
ModApté	Subonjunto de documentos da base Reuters 21578, composto de 9.603 documentos de treinamento e 3.299 documentos de teste
DF	Document Frequency
IG	Information Gain
MI	Mutual Information
CHI	CHI-squared statistic

Lista de tabelas

Lista de figuras