Ian Monteiro Nunes

Agrupamento de Registros Textuais Baseado em Similaridade Entre Textos

Dissertação de Mestrado

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

Orientador: Prof. Ruy Luiz Milidiú

Rio de Janeiro

Setembro de 2008

Ian Monteiro Nunes

Agrupamento de Registros Textuais

Baseado em Similaridade Entre Textos

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico e Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Ruy Luiz Milidiú
Orientador
Departamento de Informática – PUC-Rio

Prof. Marco Antonio Casanova
Departamento de Informática – PUC-Rio

Prof. Rubens Nascimento Melo
Departamento de Informática – PUC-Rio

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

Rio de janeiro, 04 de abril de 2008

Ian Monteiro Nunes

Graduou-se em Engenharia de Computação na PUC-Rio em 2003. Desenvolveu diversos sistemas para o meio corporativo, sendo hoje usados por empresas de abrangência nacional. Responsável pela área de pesquisa e desenvolvimento da empresa Dínamo DM desde 2004.

Ficha catalográfica

Nunes, Ian Monteiro

Agrupamento de registros textuais baseado em similaridade entre textos / Ian Monteiro Nunes ; orientador: Ruy Luiz Milidiú. – 2008.

69 f.; 30 cm

Dissertação (Mestrado em Informática)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

Inclui bibliografia

1. Informática – Teses. 2. Aprendizado de máquina. 3. Mineração de textos. 4. Deduplicação. 5. Recuperação de informação. I. Milidiú, Ruy Luiz. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III Título.

CDD: 004

Agradecimentos

Aos meus pais, Ivan e Isabel, por todo carinho, atenção, amor, apoio, paciência e tudo mais ao longo de toda a minha vida.

Ao meu orientador professor Ruy Luiz Milidiú pela parceria, disposição e ensinamentos em todas as etapas deste trabalho.

Um agradecimento especial ao meu amigo Julio Brafman por toda sua colaboração no desenvolvimento deste trabalho.

Aos meus amigos Christian Nunes, Lucas Sigaud e Eduardo Gouveia por todas as discussões e ajuda dispensada a este trabalho.

Aos professores que prontamente aceitaram participar da comissão examinadora.

À PUC-Rio pelo auxílio e estrutura fornecidos, sem os quais este trabalho não poderia ter sido realizado.

Ao professor Marcus Poggi por toda a orientação e ensinamentos ao longo de toda a minha vida acadêmica.

Aos professores e funcionários do Departamento de Informática.

A todos os amigos que direta ou indiretamente participaram deste trabalho, mesmo que não tenham sido citados nominalmente, sou igualmente grato.

Resumo

Nunes, Ian Monteiro; Milidiú, Ruy Luiz. **Agrupamento de Registros Textuais Baseado em Similaridade Entre Textos**. Rio de Janeiro, 2008, 69p. Dissertação de Mestrado. Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro

O presente trabalho apresenta os resultados que obtivemos com a aplicação de grande número de modelos e algoritmos em um determinado conjunto de experimentos de agrupamento de texto. O objetivo de tais testes é determinar quais são as melhores abordagens para processar as grandes massas de informação geradas pelas crescentes demandas de *data quality* em diversos setores da economia. O processo de deduplicação foi acelerado pela divisão dos conjuntos de dados em subconjuntos de itens similares. No melhor cenário possível, cada subconjunto tem em si todas as ocorrências duplicadas de cada registro, o que leva o nível de erro na formação de cada grupo a zero. Todavia, foi determinada uma taxa de tolerância intrínseca de 5% após o agrupamento. Os experimentos mostram que o tempo de processamento é significativamente menor e a taxa de acerto é de até 98,92%. A melhor relação entre acurácia e desempenho é obtida pela aplicação do algoritmo K-Means com um modelo baseado em trigramas.

Palavras-chave

Aprendizado de máquina; mineração de textos; deduplicação; recuperação de informação.

Abstract

Nunes, Ian Monteiro; Milidiú, Ruy Luiz. **Clustering Text Structured Data Based on Text Similarity**. Rio de Janeiro, 2008, 69p. MSc. Dissertation. Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

This document reports our findings on a set of text clustering experiments, where a wide variety of models and algorithms were applied. The objective of these experiments is to investigate which are the most feasible strategies to process large amounts of information in face of the growing demands on data quality in many fields. The process of deduplication was accelerated through the division of the data set into individual subsets of similar items. In the best case scenario, each subset must contain all duplicates of each produced register, mitigating to zero the cluster's errors. It is established, although, a tolerance of 5% after the clustering process. The experiments show that the processing time is significantly lower, showing a 98.92% precision. The best accuracy/performance relation is achieved with the K-Means Algorithm using a trigram based model.

Keywords

Machine learning; text mining; deduplicate; information retrieval.

Sumário

Índice de tabelas

Índice de figuras