Pontifícia Universidade Católica
do Rio de Janeiro

Bernardo Pereira Nunes

Classificação automática de dados semi-estruturados

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: Prof. Marco Antonio Casanova

Rio de Janeiro
abril de 2009

Pontifícia Universidade Católica
do Rio de Janeiro

Bernardo Pereira Nunes

Classificação automática de dados semi-estruturados

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.

Prof. Marco Antonio Casanova
Orientador
PUC-Rio

Prof. Antonio Luz Furtado
PUC-Rio

Profª. Karin Koogan Breitman
PUC-Rio

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

Rio de Janeiro, 03 de abril de 2009

Bernardo Pereira Nunes

Engenheiro de computação graduado pela Pontifícia Universidade Católica do Rio de Janeiro em dezembro de 2006.

Ficha Catalográfica

Nunes, Bernardo Pereira

Classificação automática de dados semi-estruturados / Bernardo Pereira Nunes ; orientador: Marco Antonio Casanova. – 2009.

92 f. : il.(color.) ; 30 cm

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

Inclui bibliografia

1. Informática – Teses. 2. Classificação. 3. Dados semi-estruturados. 4. Frames. 5. Algoritmo de classificação. 6. Classificação hierárquica. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

À minha mãe, irmãos e namorada.

Agradecimentos

Agradeço imensamente a minha querida mãe por proporcionar todas as condições para que eu pudesse realizar com êxito este trabalho, aos meus irmãos e namorada por sempre acreditarem em mim e me apoiarem ao longo da realização desta dissertação, aos amigos que possibilitaram essa caminhada mais leve e descontraída, aos amigos, em especial, Pedro Luchini e Fábio Valente por estarem sempre a disposição de ajudar e discutir as questões aqui apresentadas, a equipe CCEAD PUC-Rio pelo apoio e compreensão e, sobretudo, ao meu orientador Marco Antonio Casanova que esteve sempre presente e dedicado a realização desta obra e ao prof. Antonio Luz Furtado pela compreensão e apoio dado ao longo do curso.

Resumo

Pereira Nunes, Bernardo; Casanova, Marco Antonio. **Classificação automática de dados semi-estruturados.** Rio de Janeiro, 2009. 92p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

O problema da classificação de dados remonta à criação de taxonomias visando cobrir áreas do conhecimento. Com o surgimento da Web, o volume de dados disponíveis aumentou várias ordens de magnitude, tornando praticamente impossível a organização de dados manualmente. Esta dissertação tem por objetivo organizar dados semi-estruturados, representados por frames, sem uma estrutura de classes prévia. A dissertação apresenta um algoritmo, baseado no K-Medóide, capaz de organizar um conjunto de frames em classes, estruturadas sob forma de uma hierarquia estrita. A classificação dos frames é feita a partir de um critério de proximidade que leva em conta os atributos e valores que cada frame possui.

Palavras-chave

Classificação; dados semi-estruturados; frames; algoritmo de classificação; classificação hierárquica.

Abstract

Pereira Nunes, Bernardo; Casanova, Marco Antonio (Advisor). **Automatic classification of semi-structured data**. Rio de Janeiro, 2009. 92p. MSc. Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The problem of data classification goes back to the definition of taxonomies covering knowledge areas. With the advent of the Web, the amount of data available has increased several orders of magnitude, making manual data classification impossible. This dissertation proposes a method to automatically classify semi-structured data, represented by frames, without any previous knowledge about structured classes. The dissertation introduces an algorithm, based on K-Medoid, capable of organizing a set of frames into classes, structured as a strict hierarchy. The classification of the frames is based on a closeness criterion that takes into account the attributes and their values in each frame.

Keywords

Data classification; semi-structured data; frames; clustering algorithm; hierarchical clustering.

Sumário

1 Introdução 14

2 Teoria da Classificação 15
2.1. Classificação Decimal de Dewey (CDD) 15
2.2. Classificação Decimal Universal (CDU) 17
2.3. Classificação da Biblioteca do Congresso (LCC) 18
2.4. Classificação Facetada (Colon Classification) 19
2.5. Conclusão 21

3 Técnicas de agrupamento 23
3.1. Tipos de agrupamento 23
3.2. Algoritmos de agrupamento 24
3.2.1. Agrupamento supervisionado 24
3.2.1.1. K-Nearest Neighbor 25
3.2.1.2. Classificador Naive Bayes 26
3.2.1.3. Support Vector Machine (SVM) 26
3.2.2. Agrupamento não-supervisionado 27
3.2.2.1. Algoritmos Particionados 28
3.2.2.1.1. K-Means 28
3.2.2.1.2. K-Medóide 29
3.2.2.2. Algoritmos hierárquicos 30
3.2.2.2.1. Divisão 30
3.2.2.2.2. Aglomeração 31
3.3. Medidas de similaridade e dissimilaridade 31
3.4. Métodos de ligação sobre grupos 33
3.4.1. Single Linkage Clustering Method (SLINK) 33
3.4.2. Group Average Method ou Unweighted pair-group Method using Arithmetic Averages (UPGMA) 33
3.4.3. Complete Link Clustering Method (CLINK) 33
3.4.4. Ward's Method 34
3.5. Métodos de validação de grupos 34
3.6. Tratamento de tipos de dados 35

3.6.1. Variáveis escaladas em intervalos 35
3.6.2. Variáveis booleanas 35
3.6.3. Variáveis nominais 35
3.6.4. Variáveis ordinais 35
3.6.5. Variáveis livres 36
3.7. Determinação do número de grupos 36
3.7.1. Cross Validation 36
3.7.2. Penalized likelihood estimation 36
3.7.3. Permutation tests 36
3.7.4. Resampling 37
3.7.5. Finding the knee of error curve 37
3.8. Conclusão 37

4 Classificação Automática de dados semi-estruturados 38
4.1. Fundamentos 38
4.1.1. Noção de categorização 38
4.1.2. Noção de frames 40
4.1.2.1. Definição de frames 40
4.1.2.2. Sistemas baseados em frames 41
4.2. Implementação 42
4.2.1. Entrada de dados 43
4.2.2. Determinando o número de clusters 45
4.2.3. Algoritmo de Agrupamento 46
4.2.4. Algoritmo de Especialização 48
4.2.5. Algoritmo de Generalização 49
4.2.5.1. Noções básicas 49
4.2.5.2. Medóides 51
4.2.5.3. Abstrata 52
4.2.5.4. Híbrida 54
4.2.6. Isolamento e Coesão 54
4.2.7. Métrica de similaridade 55
4.2.8. Classificação de novos elementos 58
4.3. Análise dos dados e testes realizados 59
4.3.1. Testes com dados sintéticos 59
4.3.1.1. Teste 1 – Agrupamento de dados do tipo Person 59
4.3.1.1.1. Teste 1 – Métrica de similaridade: por atributos 60

4.3.1.1.2. Teste 1 – Métrica de similaridade: por atributos e valores 61
4.3.1.1.3. Análise do teste 1 62
4.3.1.2. Teste 2 – Classificação de dados quanto ao tipo de agrupamento 62
4.3.1.2.1. Teste com agrupamento do tipo abstrato 63
4.3.1.2.1.1. Métrica de similaridade: por atributos 63
4.3.1.2.1.2. Métrica de similaridade: por atributos e valores 65
4.3.1.2.2. Teste com agrupamento do tipo medóide 67
4.3.1.2.2.1. Métrica de similaridade: por atributos 67
4.3.1.2.2.2. Métrica de similaridade: por atributos e valores 68
4.3.1.2.3. Teste com agrupamento do tipo híbrido 69
4.3.1.2.3.1. Métrica de similaridade: por atributos 70
4.3.1.2.3.2. Métrica de similaridade: por atributos e valores 70
4.3.1.3. Teste 3 – Classificação de novos objetos 71
4.3.2. Testes com dados reais 73
4.3.2.1. Teste 1 – Análise da amostra de dados do SIAE 74
4.3.2.2. Teste 2 – Análise de dados do SIAE 77
4.4. Conclusões 78

5 Conclusões e trabalhos futuros 79

6 Referências bibliográficas 81

7 APÊNDICE A – Massa de dados sintéticos 84

Lista de figuras

Figura 1 Exemplo da Classificação de Dewey. 16
Figura 2 Exemplo da LLC. 19
Figura 3 Exemplo k-NN 25
Figura 4 SVM localiza o hiperplano h, que separa as amostras de treinamento negativas e positivas com margem máxima. Os sinais circunscritos são chamados de Support Vectors. 27
Figura 5 Exemplificação do algoritmo K-Means 29
Figura 7 Dendograma do exemplo do algoritmo de aglomeração. 31
Figura 6 Objetos a serem aglomerados. 31
Figura 8 SLINK 33
Figura 9 UPGMA 33
Figura 10 CLINK 34
Figura 11 Ilustração da coesão de um grupo (à esquerda) e a separação de dois grupos (à direita) em relação a um elemento central (protótipo). 34
Figura 12 Diagrama de relacionamentos entre frames. 42
Figura 13 Clusterização Radial. 47
Figura 14 Exemplo da terceira etapa do processo de classificação automática de dados semi-estruturados. Agrupamento de dados de medóides. 50
Figura 15 Exemplo do algoritmo de especialização sobre a classe de estudantes. 50
Figura 16 Ilustração da coesão de um cluster em relação a um elemento central (protótipo). 55
Figura 17 Ilustração da separação de dois clusters em relação a um elemento central (protótipo). 55
Figura 18 Configuração a ferramenta taxonomy creator. 59
Figura 19 Teste de classificação com objetos do tipo *Person*. 61
Figura 20 Detalhes do medóide Person_D. 62
Figura 21 Configurações de ajuste das variáveis para o processo de classificação automática. 63
Figura 22 Resultado da classificação para a massa de testes sintéticos. 63
Figura 23 Detalhes do medóide abstrato do teste de agrupamento abstrato. 64

Figura 24 Detalhes do medóide Person_C do teste de agrupamento do tipo abstrato. 65
Figura 25 Detalhes do medóide Student_G do teste de agrupamento do tipo abstrato. 65
Figura 26 Classificação utilizando a métrica por atributos e valores no agrupamento Abstrato. 65
Figura 27 Resultado da classificação utilizando a métrica de similaridade por atributos e o agrupamento por medóide. 67
Figura 28 Resultado da classificação utilizando a métrica de similaridade por atributos e valores e o agrupamento por medóide. 68
Figura 29 Resultado da classificação utilizando a métrica de similaridade por atributos e valores e o agrupamento híbrido. 70
Figura 30 Resultado do processo de classificação. 71
Figura 31 Classificação de novo objeto Person_E. 72
Figura 32 Employee_G inserido a partir da métrica ordem lexiográfica. 72
Figura 33 Comparação da classificação original do SIAE (esquerda) e da classificação gerada automaticamente (direita) pelo processo automático de classificação proposto. 77

Lista de tabelas

Tabela 1 As dez classes principais da classificação decimal de Dewey 16
Tabela 2 Classes da Biblioteca do Congresso (LLC). 19
Tabela 3 Exemplo de relacionamentos entre frames. 41
Tabela 4 Tabela do resultado da classificação para o agrupamento abstrato. 66
Tabela 5 Tabela do resultado da classificação para o agrupamento por medóide. 69
Tabela 6 Resultado do nível básico do agrupamento do teste com amostra dados reais. 75
Tabela 7 Subcategorias do medóide (f) 75
Tabela 8 Subcategorias do medóide (f.a) 76