UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
Departamento de Ciência da Computação

Plano de estudos
MAC5701 – Tópicos em Ciência da Computação

Aluno: Andre Rodrigo Sanches
Orientador: Profa. Dr. Nina S. T. Hirata
Área de Concentração: Data Mining

Tema: "Uma visão geral sobre Data Mining"

São Paulo – Outubro de 2003

1 Introdução

Um dos grandes desafios do mundo moderno é efetuar análise em grandes volumes de dados e obter conhecimento a partir deles. Alguns exemplos onde este tipo de desafio está presente são: conhecermos melhor o perfil de um cliente, descobrir associações entre os clientes e suas compras, classificarmos e categorizarmos determinados genes, buscarmos padrões similares de dados multimídia (som e vídeo), analisarmos os caminhos que usuários fazem quando navegam na internet, etc..

Na Ciência da Computação, em particular a área de Banco de Dados, não é diferente: queremos oferecer suporte eficaz, preciso, rápido, atualizável e aplicável para que os interesses deste mundo moderno se façam compreendidos e satisfeitos.

Dentro da área de banco de dados, existe a área de *data-mining* que se preocupa exatamente com este tipo de questão. Este trabalho irá focar em fornecer uma visão abrangente desta área, desde um breve histórico, os desafios e problemas encontrados, até técnicas, modelos e algoritmos mais utilizados.

2 Objetivos

O objetivo deste tabalho será adquirir conhecimentos básicos e uma visão geral da área de *data mining* através do estudo de artigos e livros que contemplam este tema. Algumas referências iniciais estão listadas ao final deste texto.

Na monografia a ser entregue no final do curso, pretendemos dar um panorama geral da área de *data mining*, desde o passado à atualidade. Relataremos os desafios e problemas encontrados tais como: diversidade de tipos de dados dos atributos (strings, inteiros, bit, texto, imagem) e tipos complexos de dados (som, animações, multimídia, vídeos, etc), grandes volumes de dados (gigabytes, terabytes), disponibilidade e atualização/carga dos sistemas e informações.

Abordaremos as técnicas mais utilizadas: regras de classificação; generelização, sumarização e OLAP; classificação de dados; aglomeração (*clustering*); busca de padrões e similaridade; padrões de caminhos (*path traversal pattern*). Faremos indicações no uso das técnicas e exemplos de aplicação.

Descreveremos os modelos e algoritmos utilizados: redes neurais; árvores de decisão; MARS/MBR; indução de regras/regressão; algoritmos genéticos.

Uma técnica muito utilizada atualmente em data-mining é a aglomeração. Por isso, daremos ênfase em 4 algoritmos muito conhecidos (utilizados, principais) de aglomeração em bancos de dados : CLARANS, BIRCH, DBSCAN e CURE.

3 Resultados esperados

Além da monografia em si, esperamos que os conhecimentos a serem adquiridos durante a elaboração da mesma sirvam como subsídios diretos no desenvolvimento da dissertação de mestrado.

Referências

[1] Ming-Syan Chen, Jiawei Han, Philip S. Yu *Data Mining: An Overview from Database Perspective.*

[2] Two Crows Corporation *Introduction to Data Mining and Knowledge Discovery.*

[3] Dunham MH. *Introductory and Advanced Topics. Prentice-Hall. 2002.*

[4] Han J, Kamber M. *Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.*

[5] Krzysztof Koperi, Junas Adhikary, Jiawei Han *Spatial Data Mining: Progress and Challenges Survey paper.*

[6] J. Han, M. Kamber, A. K. H. Tung *Spatial Clustering Methods in Data Mining: A Survey.*

[7] Raymond T. Ng, Jiawei Han *CLARANS: A Method for Clustering Objects for Spatial Data Mining.*

[8] Tian Zhang, Raghu Ramakrishnan, Miron Livny *BIRCH: An Efficient Data Clustering Method for Very Large.*

[9] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim *CURE: An Efficient Clustering Algorithm for Large Databases.*

[10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.*