Descoberta de Conhecimento em Bancos de Dados

Marco Eugênio Madeira Di Beneditto
Orientador: Prof Leliane Nunes de Barros

1 Proposta de Estudo

Esta proposta foca a descoberta de conhecimento em banco de dados utilizando-se de conhecimento prévio sobre os dados. O objetivo deste estudo é apresentar diferentes enfoques e implementações, analisando suas utilizações e elaborar um programa para testes. Este estudo faz parte da pesquisa de mestrado realizada pelo autor.

2 Motivação

A maioria dos estudos em descoberta de conhecimento somente procura por padrões no nível conceitual representado na base de dados que, nem sempre, é o nível de conhecimento que se deseja ou mesmo o nível que evidencia conhecimento de interesse. Além disso, dados com grande regularidade, i.e., grande ocorrência numa base de dados, usualmente não aparecem em níveis conceituais baixos pois, neste caso, os padrões tendem a ser dispersos. Como exemplo, têm-se os registros das compras efetuadas num supermercado onde apenas os códigos de barra são registrados e, dificilmente descobrem-se os padrões que associam os itens entre si numa mesma compra, como pão integral e leite desnatado ao invés do pão marca X e leite marca Y. Por outro lado, dados num nível conceitual mais elevado, mais geral, podem levar a padrões que já são de conhecimento comum, não possuindo algo de novo.

Estas observações motivaram a proposta de descobrir conhecimento em múltiplos níveis conceituais [Agrawal & Srikant, 1995], [Fu, 1996] e [Taylor, 1999], aumentando a diversidade dos padrões descobertos e permitindo ao usuário focar em pontos que lhe sejam mais interessantes.

Um padrão a ser descoberto pode ser representado por uma expressão, em alguma linguagem, descrevendo um subconjunto dos dados ou um modelo aplicável a este subconjunto. Os seguintes padrões podem ser descobertos num nível conceitual mais elevado:

Classificação classificação é uma função de aprendizado que mapeia dados de entrada em um número finito de classes pré-definidas, podendo ser representada através de regras.

Regras de Caracterização descrevem as regras de maior relevância que caracterizam determinada classe, onde o conjunto de dados analisados pertencem a apenas uma classe.

Regras de Discriminação descrevem as regras que melhor discriminam os dados de uma classe alvo em relação a uma outra classe, denominada classe contrastante, onde o conjunto de dados são exemplos das duas classes.

Associação caracterizam o quanto a presença de um conjunto de atributos nos registros de um banco de dados implica na presença de um outro conjunto distinto de atributos nos mesmos registros.

Generalização e Sumariação envolvem métodos de descoberta de formas compactas de descrição de um conjunto de dados. Apresentam as características gerais ou uma visão de alto nível reduzida sobre uma base de dados.

Agrupamentos ou *Clustering* agrupam um conjunto de dados, sem uma classe pré-definida, baseado no conceito de maximização da similaridade intraclasse e minimização da similaridade entre as classes.

3 Mineração em múltiplos níveis

Com a disponibilidade de uma hierarquia de conceitos, os padrões descobertos podem ser declarados envolvendo conceitos em diferentes níveis de abstração, conduzindo a regras de múltiplos níveis, o que é geralmente mais desejável. Por exemplo, na descoberta de regras de classificação do tipo: **se** A **então** C_n, onde A representa o conjunto de pares atributo-valor $\langle A_i, v_k \rangle$, utilizando-se uma hierarquia para A_i, v_k poderá assumir um dos k-valores do atributo dentre os disponíveis na hierarquia de conceitos do atributo A_i. Deste modo a regra descoberta poderá ser expressa num nível conceitual mais elevado do que aquele representado na base de dados.

A descoberta de uma regra de classificação a partir de um conjunto de exemplos positivos e negativos em relação a classe pode ser tratada como uma busca através de um espaço de descrições, i.e., regras candidatas. Se cada tupla possuir i atributos com k possíveis valores para cada um (aqui já incluídos os valores das hierarquias de conceitos de cada atributo) há k^i possíveis combinações de tuplas. Qualquer subconjunto dos atributos pode corresponder ao antecedente de uma regra o que significa 2^{k^i} regras possíveis.

São poucos os trabalhos encontrados na literatura sobre descoberta de regras de classificação em múltiplos níveis: [Han, 1996], propõem a aplicação do princípio de Indução Orientada a Atributo, para obter tabelas em níveis conceituais mais elevados e em [Fu, 1996] são utilizados os algoritmos de mineração sobre estas tabelas. Em [Taylor, 1999] é desenvolvido um algoritmo de indução de regras de classificação em múltiplos níveis que utiliza processamento paralelo para as buscas envolvidas.

A descoberta de um novo padrão gera a necessidade de registrá-lo para posterior uso numa outra tarefa de descoberta de conhecimento. Este tipo de registro, que é mais complexo que uma simples relação de generalização-especialização, suscita o estudo e emprego de outras formas de representação do conhecimento que não sejam apenas hierarquias em forma de árvore. O uso de ontologias, que possui uma maior capacidade de representação (não apenas relações de generalização-especialização), permite este tipo de armazenamento.

Este estudo é motivado pela suposição que através dessas abordagens é possível realizar a mineração de dados com o uso de conhecimento prévio não representado no banco de dados.

4 Plano de Estudos

Este plano de estudos seguirá o seguinte cronograma:

Março/Abril - Levantamento bibliográfico de trabalhos na área e resumos dos trabalhos lidos.

Maio - Análise e seleção de um ambiente de banco de dados para implementar um algoritmo de mineração de dados em múltiplos níveis.

Junho - Elaboração de um resumo final com os resultados obtidos a ser entregue como relatório desta disciplina.

Referências

[Agrawal, Imielinski & Swami, 1993] AGRAWAL, Rakesh; IMIELINSKI, Tomasz; SWAMI, Arun. *Mining Association Rules between Sets of Items in Large Databases* 1993. 10p. Proceedings of the 1993 ACM SIGMOD Conference Washington DC, EUA.

[Agrawal & Srikant, 1995] AGRAWAL, Rakesh; SRIKANT, Ramakrishnan. *Mining Generalized Association Rules.* 1995. 13p. Proceedings of 21st VLDB Conference Zurich, Suíça.

[Cai, Cercone & Han, 1992] CAI, Y; CERCONE, Nick; HAN, Jiawei. *Knowledge Discovery in Databases: An attribute-oriented approach.* 1992. 12p. Proceedings of the 18th Conference in Very large Databases, Vancouver, Canadá.

[Cercone, Hamilton & Hilderman, 1999] CERCONE, Nick; HAMILTON, Howard J.; HILDERMAN, Robert J. *Data Mining in Large Databases Using Domain Generalization Graphs.* 1999. 42p. Kluwer Academic Publishers.

[Fensel & Klein, 2001] FENSEL, Dieter; KLEIN, Michel. *Ontology versioning on the Semantic Web.* 2001. 18p. Vrije Universiteit Amsterdam, Amsterdã, Holanda.

[Fu, 1996] FU, Yongjian. *Discovery of Multiple-Level Rules from Large Databases.* 1996. 184p. PhD Thesis - Scool of Computer Science of Simon Fraser University, Burnaby, B.C., Canadá.

[Fudger & Hamilton, 1995] FUDGER, David R.; HAMILTON, Howard J. *Estimating DBLEARN's Potential for Knowledge Discovery in Databases.* 1995. 17p. Computational Intelligence, vol 11, nr. 2 Blackwell Publishers.

[Han, 1996] HAN, Jiawei. *Mining Knowledge at Multiple Concept Levels.* 1996. 6p. Invited talk - Scool of Computer Science of Simon Fraser University, British Columbia, Canadá.

[Lu, 1997] LU, Yijun. *Concept Hierarchy in Data Mining: Specification, Generation and Implementation.* 1997. 106p. MSc Thesis - Scool of Computer Science of Simon Fraser University, Burnaby, B.C., Canadá.

[Taylor, 1999] TAYLOR, Merwyn G. *Finding High Level Discriminant Rules in Parallel.* 1999. 226p. PhD Thesis - Faculty of the Graduate School of the University of Maryland, College Park, EUA.