FastMapDB: Uma Ferramenta para Visualização em SGBDRs com suporte à Filtragem e Seleção Visual dos Dados

Adriano Arantes Paterlini[1], **Rodrigo Fálico Teodoro de Faria**[1], **Humberto L. Razente**[1], **Caetano Traina Júnior**[1], **Agma Juci Machado Traina**[1]

[1]Departamento de Computação – Universidade de São Paulo
Av. do Trabalhador São Carlense, 400, Centro
Caixa Postal 668, CEP 13560-970, São Carlos - SP, Brasil

paterlini@gmail.com, {rodf@grad, hlr@, caetano@, agma@}icmc.usp.br

Abstract. *In the last decades, a huge amount of data was stored and that led to the creation of automatic data analysis techniques. Unfortunately, those techniques find problems to take advantage of all the knowledge inherent in the data. This paper introduces the FastMapDB tool which, based on the mankind high capability to interpret graphical information, allows graphic visualization of data stored into a relational base be independent of the summarization process it had been applied to. The FastMapDB allows its users to play an interactive role in the process of knowledge discovering as, for example, the utilization of filters or through functionalities as the visual data selection which allows them to select objects from the mapping and then to analyse it separately.*

Resumo. *A grande quantidade de dados armazenada nas últimas décadas fez surgir muitas técnicas automáticas para análise de dados, mas, freqüentemente, estas encontram problemas para aproveitar toda a informação oculta nos dados. Este artigo apresenta a ferramenta FastMapDB, a qual, aproveitando a alta capacidade humana para interpretar informações em formato gráfico, permite a visualização de forma gráfica dos dados armazenados em uma base relacional, independente de processos de sumarização que tenham sido aplicados. O FastMapDB fornece aos usuários uma participação interativa no processo de descoberta de conhecimento como, por exemplo, a utilização de filtros ou através de funcionalidades como a seleção visual dos dados, o que lhe permite selecionar conjuntos de objetos do mapeamento e analisá-los separadamente.*

1. Introdução

Nas últimas décadas a capacidade das empresas de gerar e coletar informações aumentou rapidamente. Esta explosão no volume de dados gerou a necessidade do desenvolvimento de novas técnicas e ferramentas que pudessem, além de processar os dados, permitir sua análise, de maneira inteligente e automática, para a descoberta de informações úteis. Isto acarretou, então, o surgimento de um proeminente campo de pesquisa para a extração de informação em bases de dados, denominado Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases - KDD) [Frawley et al., 1991]. KDD é o processo de identificar padrões ou modelos que representem informação válida, inédita, potencialmente útil e essencialmente compreensível em uma coleção de dados [Fayyad, 1998]. Em geral, técnicas de mineração de dados possuem um papel preponderante neste processo.

Devido a natureza das informações armazenadas nestes grandes volumes de dados serem textuais e numéricas, especialmente com muitos atributos, a interpretação destes é realizada com baixa eficiência por seres humanos. Apresentá-los de forma gráfica permite

uma melhor percepção dos dados e que informações possam ser exploradas e extraídas a partir das visualizações criadas. Através da visualização espacial pode-se, por exemplo, verificar a existência de casos de exceção (*outliers*) e a formação de agrupamentos (*clusters*).

Este artigo apresenta uma ferramenta para exploração visual de dados denominada *FastMapDB* (FastMap in Data Bases) [Barioni et al., 2002, Traina et al., 2001] que mostrou-se útil para análise de dados em KDD. Está dividido da seguinte forma: a Seção 2 apresenta uma descrição da ferramenta *FastMapDB*, na Seção 3 é mostrado com mais enfoque a seleção visual dos dados e na Seção 4 são apresentados os resultados.

2. A Ferramenta *FastMapDB*

O núcleo da ferramenta *FastMapDB* utiliza o algoritmo denominado *FastMap* [Faloutsos and Lin, 1995] que realiza o mapeamento de objetos, que podem estar em espaços de diferentes dimensões, para o espaço Euclidiano. O mapeamento feito com o *FastMap* procura preservar as distâncias entre os dados e minimizar possíveis distorções do mapeamento.

A ferramenta *FastMapDB* considera que uma relação em uma base de dados relacional é uma coleção de pontos em um espaço N-dimensional (número de atributos não-chave da relação) na qual cada atributo representa um valor numérico (medidas, valores monetários etc), uma data, ou um texto curto (nomes). A cada um destes domínios é definida uma função distância (módulo da diferença entre dois valores numéricos; número de dias entre duas datas; ou distância de edição entre dois textos). As distâncias individuais entre cada atributo são então combinadas para definir a distância entre cada tupla da relação, o que permite uma tupla de N atributos ser interpretada como um ponto em um espaço N-dimensional. Assim, o algoritmo de redução de dimensionalidade (extração de atributos) pode mapear esse espaço N-dimensional para um espaço tridimensional no qual a visualização dos dados é realizada. Isto permite a utilização da habilidade humana de interpretação visual para analisar os dados armazenados e descobrir (minerar visualmente) a informação “oculta”.

Outra característica é permitir ao analista compreender como os itens de dados estão relacionados, quais são os padrões envolvidos e detectar agrupamentos e elementos de exceção. A visualização gerada pela ferramenta é simples e rapidamente manipulada, o que permite que hipóteses sejam formuladas e testadas antes que se efetue cálculos estatísticos e numéricos computacionalmente caros.

Há, também, recursos para auxiliar o usuário na escolha de conjuntos reduzidos de atributos o que tornam as tarefas de mineração mais eficientes e que permitem a sobreposição de diferentes mapeamentos de um mesmo conjunto de dados de alta dimensão. Igualmente, torna possível a observação visual das correlações existentes entre os atributos ou entre amostras de dados obtidas em diferentes instantes ou com diferentes parametrizações [Razente et al., 2004].

A ferramenta permite ainda utilizar um atributo da relação (com domínio discreto) para classificar as tuplas, possibilitando a visualização de classes em diversas cores e formatos, além de permitir a utilização de diversos filtros para selecionar apenas as tuplas de interesse.

3. Utilização da seleção visual dos dados

A ferramenta *FastMapDB*, disponível para download em http://www.gbdi.icmc.usp.br, foi desenvolvida utilizando a linguagem C++ e a biblioteca gráfica OpenGL

(http://www.opengl.org) para manipulação visual dos dados, o que permite ao usuário uma interação intuitiva. Além disso, foi desenvolvida para operar sobre quaisquer SGBDRs que suportem ODBC (Open Database Connectivity).

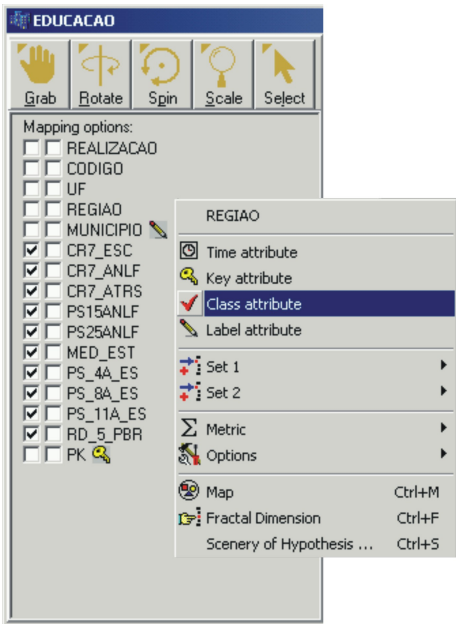

Figura 1: Escolha dos atributos a serem mapeados e definição dos parâmetros da visualização

Inicialmente o usuário escolhe a relação $\Re$ e os atributos desejados. É permitido, também, designar alguns parâmetros específicos aos atributos, como mostra a Figura 1, tais como *time attribute*, que separa os mapeamentos por período; *key attribute*, que define o atributo como chave do mapeamento; *class attribute* que é um atributo categórico; e *label attribute* que é um identificador para objetos a serem mapeados. Além disso, é possível escolher a métrica a ser aplicada (por padrão, a métrica é a L2) através da opção *Metric* na lista de opções mostrada na Figura 1. As possíveis métricas são da família Minkowski, entre elas estão $L\infty$ (*L Infinity* ou *Chebychev*), L1 (*Manhatan*), L2 (*Euclidean*) ou, ainda, qualquer outra de potência maior que um.

Uma vez estabelecidos os parâmetros e métricas desejados para a visualização, o mapeamento pode ser realizado através da escolha da opção *Map* da lista de opções mostrada na Figura 1. A Figura 2 mostra a visualização de um mapeamento utilizando a base de dados IPEA (http://www.ipeadata.gov.br) da relação Educação, que são índices relacionados ao ensino nas cidades brasileiras, e a métrica L2. O mapeamento realizado é classificado pelas regiões e identificado pelos nomes das cidades.

A visualização é altamente interativa. Há diversos recursos para a manipulação do mapeamento, o que auxilia na visualização tridimensional e na recuperação de informações, sendo eles *Grab*, que permite deslocar o mapeamento; *Rotate*, que rotaciona o mapeamento; *Spin*, que gira o mapeamento; e *Scale*, que muda o tamanho da visualização. Além das funcionalidades descritas, há também a funcionalidade *Select*, a qual auxilia na identificação dos objetos mapeados pois, através da utilização de formas geométricas especificas, permite a seleção dos objetos desejados na visualização. Este recurso possui como opções:

- A seleção múltipla (*Multiple Selection*), que permite a utilização de mais de uma forma geométrica na mesma visualização;
- *Plane*, que seleciona os objetos utilizando um plano divisor;
- *Point*, que faz a seleção pontual dos objetos;

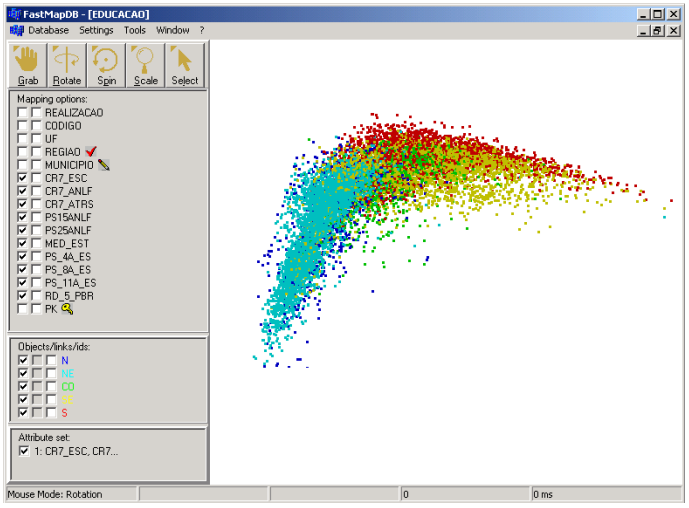

Figura 2: Visualização exemplo do mapeamento da relação Educação

- *Sphere*, que seleciona os objetos utilizando uma esfera como forma geométrica;
- *Square*, que seleciona os objetos utilizando um paralelepípedo como forma geométrica;
- *Move*, que permite ao usuário mover a seleção realizada;
- *Scale*, que permite redimensionar a seleção;
- *Reset*, que elimina todas as seleções do mapeamento;
- *Invert*, que inverte a atual seleção, isto é, os objetos atualmente selecionados perdem sua seleção e os outros são selecionados; e
- *Inverse Selection*, que mostra o resultado da consulta visual realizada.

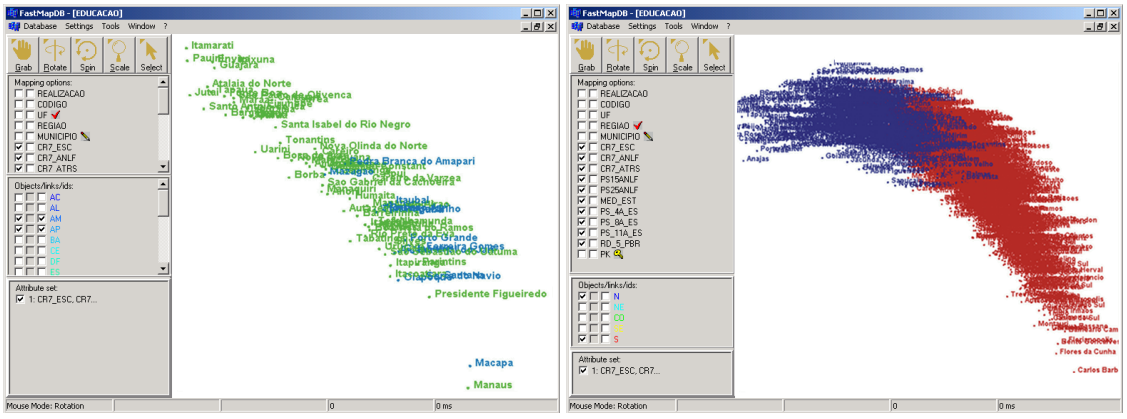

Figura 3: Visualização resultado da aplicação do filtro em que são mostrados os nomes das cidades (a) Objetos referentes aos Estados do Amazonas (AM) e do Amapá (AP). (b) Objetos referentes às Regiões Sul e Norte

A utilização dos filtros disponíveis permite que, por exemplo, utilizando-se da seleção de classes, na qual o grupo classificado pode ser visualizado separadamente, apenas os objetos referentes às cidades dos Estados do Amazonas (AM) e do Amapá (AP) sejam mostrados, assim como na Figura 3 (a). Nesta visualização ainda é possível (interagindo com os recursos de escala e rotação) identificar as cidades, pois há poucos objetos. Entretanto, como pode ser visto na visualização da Figura 3 (b), na qual são mostrados os objetos referentes às cidades da Região Sul e Norte, há certas situações em que é difícil identificar os objetos devido a seu grande número.

Já com a utilização da seleção visual, é possível contornar o problema apresentado na Figura 3 (b) pois, através das formas de seleção, pode-se recuperar apenas as

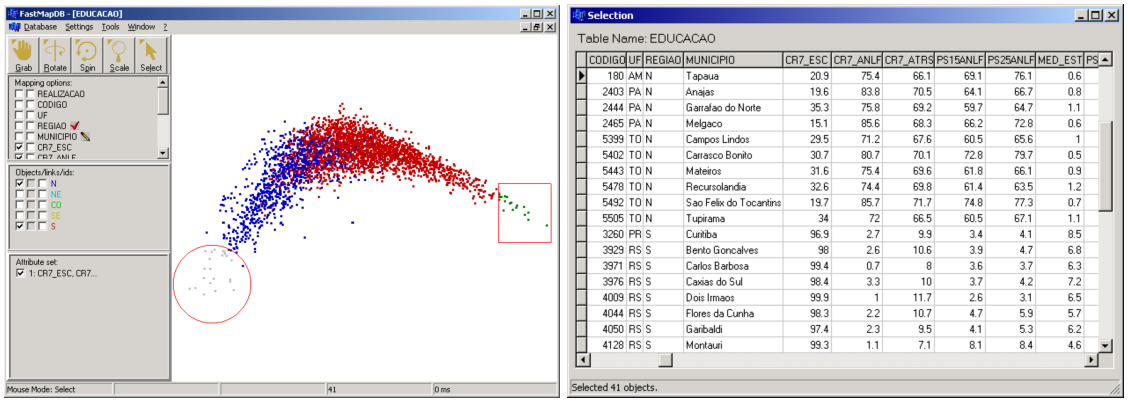

Figura 4: (a) Visualização dos objetos com múltiplas seleções referentes às Regiões Norte e Sul (b) Resultado do mapeamento inverso desta seleção (*Inverse Selection*)

informações dos objetos selecionados na visualização, isto é, realiza-se uma consulta visual da relação mapeada. As Figuras 4 e 5 exemplificam uma utilização do recurso. Na visualização da Figura 4 (a) foram utilizadas múltiplas seleções, uma na forma de esfera e outra na forma de paralelepípedo, para selecionar os dois extremos da visualização. No extremo esquerdo foram selecionados os objetos referentes às cidades com menores índices de educação da Região Norte, enquanto que os objetos selecionados no extremo direito são referentes às cidades com os melhores índices de educação da Região Sul. Estas descobertas foram obtidas por meio do uso deste recurso que permite a análise parcial dos dados. Os objetos selecionados são mostrados como resultado de uma consulta na relação da Figura 4 (b).

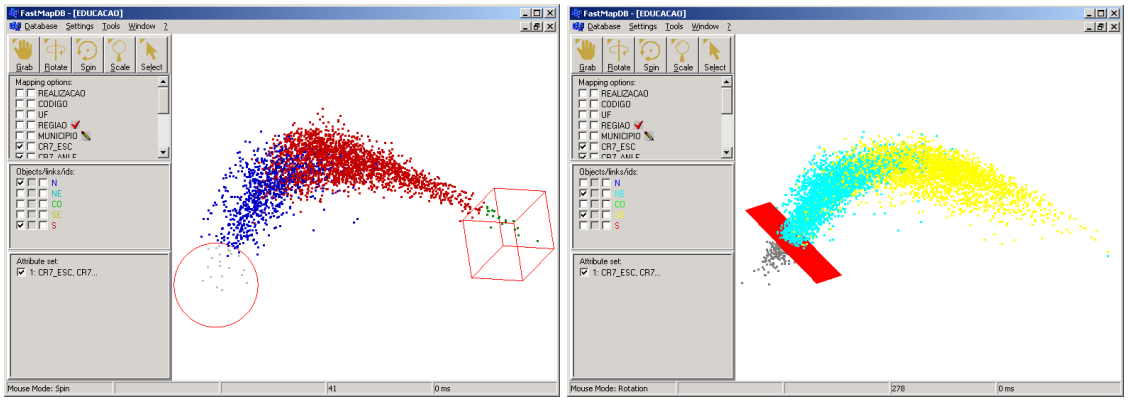

Figura 5: Visualizações rotacionadas dos objetos referentes às (a) Regiões Norte e Sul com aplicação de múltiplas seleções (b) Regiões Nordeste e Sudeste com aplicação da seleção em forma de plano

Da mesma forma, a Figura 5 (b) mostra uma visualização na qual foi utilizada a seleção em forma de plano para obter os objetos referentes às Regiões Nordeste e Sudeste. Nesta forma de seleção os objetos selecionados são aqueles "abaixo" do plano, isto é, aqueles que se encontram na direção oposta ao vetor normal à superfície definida. Em ambos os itens da Figura 5 trata-se da aplicação do recurso de rotação sobre os objetos já selecionados. Esta interação entre os recursos da ferramenta *FastMapDB* favorece a análise dos dados pois possibilita uma melhor visualização e posterior refinamento da seleção.

Além das varias funcionalidades disponiveis a ferramenta *FastMapDB* tem se mostrado escalável. Testes realizados com microcomputador com o processador AMD Athlon XP 2400+, 512 MB de memória RAM e Windows 2000. A escalabilidade da

ferramenta quanto ao número de atributos selecionados para o mapeamento, variando o número de atributos em dois conjuntos de dados, com 50.000 e 100.000 objetos respectivamente, compostos por 24 atributos uniformemente distribuídos aleatoriamente. Para o conjunto com 100.000 objetos, o tempo gasto com o mapeamento de 12 atributos para 3 dimensões foi de aproximadamente 2 segundos, enquanto que o tempo gasto com o mapeamento de 24 atributos para 3 dimensões foi inferior a 4 segundos.

4. Conclusão

A visualização gráfica dos dados utilizando a ferramenta *FastMapDB* permite que a capacidade humana seja utilizada na interpretação dos dados possibilitando, portanto, a inclusão do usuário como parte importante no processo de análise dos dados para a descoberta de conhecimento em grandes bases de dados.

Ao contrário das ferramentas para visualização científica, o *FastMapDB* é uma ferramenta para visualização de informações que cria uma distribuição para os dados armazenados em uma relação de uma base de dados relacional quaisquer que sejam os tipos de dados dos atributos que a constituem não sendo necessário que a mesma tenha uma dimensão espacial intrínseca.

A seleção visual dos dados é um dos recursos presentes no *FastMapDB*, que amplia a capacidade da ferramenta auxiliando o usuário a recuperar informações da visualização mapeada a partir das relações contidas nas bases de dados. Isto é feito possibilitando-se a recuperação das tuplas mapeadas por meio de uma consulta visual baseada nas seleções criadas interativamente utilizando formas geométricas e a seleção pontual de objetos.

Os recursos disponíveis no *FastMapDB* e a sua escalabilidade possibilitam, também que a ferramenta seja utilizada na validação de algoritmos e análises de dados, tornando o processo de mineração de dados mais interativo. Isto é válido, pois permite verificar se processos tais como agrupamento, classificação etc, obtiveram resultados positivos ou ajudar na correção de problemas.

Referências

Barioni, M. C. N., Botelho, E., Faloutsos, C., Razente, H. L., Traina, A. J. M., and C. Traina, J. (2002). Data Visualization in RDBMS. In *IASTED Intl. Conf. Information Systems and Databases (ISDB)*, Tokyo, Japan.

Faloutsos, C. and Lin, K.-I. D. (1995). FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. In *ACM Int'l Conf. on Data Management (SIGMOD)*, p. 163–174, Zurich, Switzerland. M. Kaufmann.

Fayyad, U. M. (1998). Mining databases: Towards algorithms for knowledge discovery. *Bullettin of Tech. Committee on Data Engineering*, 21(1):29–48.

Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1991). Knowledge Discovery in Databases: An Overview. p. 1–27. AAAI/MIT Press.

Razente, H. L., Chino, F. J. T., Barioni, M. C. N., Traina, A. J. M., and Traina, Caetano, J. (2004). Visual analysis of feature selection for data mining processes. In Lifschitz, S., editor, *Simpósio Brasileiro de Banco de Dados*, vol. 1, p. 163–177, Brasília, DF. SBC.

Traina, A. J. M., Traina, Caetano, J., Barioni, M. C. N., Botelho, E., and Bueno, R. (2001). Visualização de dados em sistemas de bancos de dados relacionais. In *XVI Simpósio Brasileiro de Banco de Dados*, p. 95–109, Rio de Janeiro.