Desenvolvimento de um Framework para Análise Visual de Informações Suportando Data Mining

José Fernando Rodrigues Júnior

Orientadora: *Profa. Dra. Agma Juci Machado Traina*

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP como parte dos requisitos para obtenção do título de Mestre em Ciências de Computação e Matemática Computacional.

USP – São Carlos
Julho de 2003

Este trabalho foi realizado com apoio financeiro da Fapesp - <http://www.fapesp.br>,
processo número 01/11287-1.

Dedico este trabalho à minha Família, um motivo constante de alegrias, à minha noiva Fabíola, que me motiva em querer sempre mais, e à minha orientadora Agma, sempre atenciosa e fundamental para minha formação.

Índice

Capítulo 1 - Introdução	1
1.1 - Considerações Gerais	1
1.2 - Motivação e Objetivos	2
1.3 - Apresentação e Organização do Trabalho	4
Capítulo 2 - Descoberta de Conhecimento em Bases de Dados	6
2.1 - Considerações Gerais	6
2.2 - As Etapas do KDD	7
2.3 - Mineração de Dados	8
2.3.1 - Principais Abordagens da Mineração de Dados	9
2.3.2 - Uma Proposta de Metodologia para Aplicação da Mineração de Dados	10
2.3.3 - Arquiteturas de Sistemas de Mineração de Dados	12
2.4 - Implicações sobre um Sistema de Informação	13
2.5 - Visualização, Mineração de Dados e Mineração Visual de Dados	15
2.6 - Considerações Finais	18
Capítulo 3 - Visualização de Informações	19
3.1 - Considerações Gerais	19
3.2 - Visualização de Informações x Visualização Científica	20
3.3 - Técnicas de Pré-processamento dos Dados	21
3.4 - Interação	23
3.4.1 - Técnicas de Interação	23
3.4.2 - Um Modelo de Interação	26
3.5 - Técnicas de Visualização	28
3.6 - Exemplos de Técnicas de Visualização	32
3.7 - Desempenho no projeto de técnicas de visualização	40
3.8 - Limitações das técnicas de visualização	43
3.9 - Considerações Finais	44
Capítulo 4 - O Projeto Desenvolvido	45
4.1 - Considerações Gerais	45
4.2 - A Ferramenta FastMapDB	45
4.3 - Duas contribuições à identificação visual de aglomerados na ferramenta FastMapDB	48
4.4 - Técnicas de Visualização de Informações Utilizadas	52
4.5 - Integração das Técnicas	54
4.6 - Exibição Visual de Frequências	56
4.7 - Exibição de Dados por Relevância	59
4.8 - Exibição Visual de Dados Estatísticos	64
4.9 - Características do Projeto de Software	67
4.10 - Arquitetura de Desenvolvimento	68
4.11 - O <i>Pipeline</i> de Visualização	71
4.12 - Considerações Finais	74
Capítulo 5 - Conclusões e Linhas de Futuras Pesquisas	75
5.1 - Considerações Finais	75
5.2 - Sugestões de Futuras Pesquisas	76
Referências Bibliográficas	77

Lista de Figuras

Figura 1 -	6
As etapas que constituem o processo de KDD (extraído de (Fayyad, Piatetsky-Shapiro et al. 1996)).	
Figura 2 -	25
Modelo mecânico (a) que caracteriza as técnicas <i>Perspective Wall</i> e <i>Bifocal Displays</i> . Em (b) é apresentada a aparência do espaço de dados transformado pela técnica ilustrada em (a). Extraído de (Leung and Apperley 1994).	
Figura 3 -	28
Modelo de interação extraído de (Keim, Lee et al. 1995). (a) A arquitetura atual tendo o usuário como componente central. (b) A arquitetura proposta tendo a visualização como componente central.	
Figura 4 -	30
(a) Técnica de visualização orientada a <i>pixels</i> baseada em uma consulta sobre uma base de dados de cinco dimensões. (b) Uma alternativa de arranjo para apresentação de todos os atributos em uma única janela, como visto em (a).	
Figura 5 -	33
Coordenadas Paralelas, na cena uma filtragem interativa do conjunto de dados Carros de origem japonesa (verde) e europeia (azul) que possuem quatro cilindros. Gerado com auxílio da ferramenta GBDIView.	
Figura 6 -	35
Scatter Plots com <i>Link & Brush</i> : a base de dados de carros exibida com a mesma seleção realizada na visualização das Coordenadas Paralelas na figura 5. Em destaque a relação "milhas por galão x peso" dos carros japoneses (verde) e europeus (azul) que possuem quatro cilindros. Gerado com auxílio da ferramenta GBDIView.	
Figura 7 -	36
Visualização de dados estatísticos do conjunto Iris através do <i>StarCoordinates</i> sem os pontos (a), e com os pontos (b). O eixo da dimensão <i>class</i> não está ativado em nenhuma das cenas, aqui ele é utilizado apenas para realização de filtragem interativa, sendo que vermelho representa a classe versicolor, verde a classe virginica e azul a classe setosa. Em (c) é descrito o cálculo da posição dos pontos no plano cartesiano. Ilustração gerada com auxílio da ferramenta GBDIView.	
Figura 8 -	37
A técnica <i>Table Lens</i> , agora, nesta visualização apresenta a seleção dos veículos norte-americanos (azul) e a seleção dos veículos europeus (verde) com quatro ou oito cilindros. Os dados estão ordenados por número de cilindros, o que deixa clara a correlação entre este atributo e os demais: a autonomia, a potência, o peso, a aceleração e até mesmo o ano de fabricação e a origem. Carros mais novos e europeus tendem a ter um número menor de cilindros. Gerado com auxílio da ferramenta GBDIView.	
Figura 9 -	38
<i>Star Glyphs</i> exibindo a base de flores da espécie Iris. Em vermelho temos o gênero setosa. Visualização gerada com auxílio da ferramenta XMDV (Ward 1994).	
Figura 10 -	39
Esquema hierárquico da técnica <i>Dimensional Stacking</i> , gerada com auxílio da ferramenta XMDV (Ward 1994), o destaque em vermelho representa o gênero setosa da espécie Iris que possui sépalas mais largas e estreitas, e pétalas menores que as demais. Em verde temos os gêneros versicolor e virginica.	

Figura 11 -	46
O FastMapDB e seus vários elementos de interação e análise.	
Figura 12 -	51
Em (a) temos exemplares dos sete grupos de imagens utilizadas no experimento. (b) mostra a curva de <i>Precision and Recall</i> dos quatro extratores ao responder consultas por similaridade (vizinhos mais próximos). Em (c) temos a visualização dos vetores de características extraídos das imagens exemplificadas em (a). (d) mostra as métricas calculadas pela ferramenta FastMapDB.	
Figura 13 -	55
Os selecionadores utilizados para interagir com os componentes visuais. (a) Coordenadas Paralelas, (b) <i>Scatter Plots</i> , (c) Coordenadas Estelares, e (d) <i>Table Lens</i> .	
Figura 14 -	57
Ilustração da técnica <i>Frequency Plot</i> sobre a base de dados de exames laboratoriais de tecidos para identificação de câncer. Em (a) é exibida a totalidade da base de dados segundo a frequência dos valores que a compõe. Em (b) e (c), respectivamente, podem ser observados os tecidos sadios (classe 0) e os tecidos doentes (classe 1) segundo a filtragem interativa simples. Em (d) e (e) as mesmas seleções utilizando-se da técnica <i>Frequency Plot</i> . Uma breve análise das imagens (d) e (e) possibilita a caracterização dos tecidos sadios e dos tecidos doentes.	
Figura 15 -	62
Exemplo de cálculo do <i>DRC</i> para um registro sendo exibido em uma cena das Coordenadas Paralelas.	
Figura 16 -	63
Ilustração do <i>Relevance Plot</i> . Na cena temos a base de dados veículos sem o atributo de classe. Através da seleção especulativa proporcionada pela técnica é possível a identificação de duas tendências: em (a) veículos mais econômicos, menos potentes, mais leves e rápidos; em (b) veículos com menor desempenho e aceleração, mais potentes e pesados.	
Figura 17 -	65
Exemplo de exibição de dados estatísticos sobre cena de visualização. Na imagem de uma cena da técnica de Coordenadas Estelares temos as médias em verde, os valores de moda em rosa, os desvios padrões em azul claro e as medianas em azul escuro. É mostrado também o menu de interação oferecido ao usuário.	
Figura 18 -	69
A arquitetura dos componentes desenvolvidos. Os componentes de hardware são apresentados em cinza. Os componentes de software de terceiros são apresentados em amarelo. Em azul estão indicados os softwares desenvolvidos, ou que requerem implementação para que uma nova técnica seja desenvolvida.	
Figura 19 -	73
Exemplo de utilização do <i>Pipeline</i> de visualização. Em (a) a visualização global dos dados referentes às cidades; os detalhes verdes indicam os agrupamentos selecionados para visualização multimodal. Em (b) temos a visualização, por Coordenadas Estelares, do agrupamento menor; em (c) a visualização do maior.	

Lista de Tabelas

Tabela 1 - As técnicas utilizadas no projeto e suas características.	53
---	----

Resumo

No presente documento são reunidas as colaborações de inúmeros trabalhos das áreas de Bancos de Dados, Descoberta de Conhecimento em Bases de Dados, Mineração de Dados, e Visualização de Informações Auxiliada por Computador que, juntos, estruturam o tema de pesquisa e trabalho da dissertação de Mestrado: a Visualização de Informações. A teoria relevante é revista e relacionada para dar suporte às atividades conclusivas teóricas e práticas relatadas no trabalho.

O referido trabalho, embasado pela substância teórica pesquisada, faz diversas contribuições à ciência em voga, a Visualização de Informações, apresentando-as através de propostas formalizadas no decorrer deste texto e através de resultados práticos na forma de softwares habilitados à exploração visual de informações. As idéias apresentadas se baseiam na exibição visual de análises numéricas estatísticas básicas, frequenciais (*Frequency Plot*), e de relevância (*Relevance Plot*). São relatadas também as contribuições à ferramenta FastMapDB do Grupo de Bases de Dados e Imagens do ICMC-USP em conjunto com os resultados de sua utilização. Ainda, é apresentado o Arcabouço, previsto no projeto original, para construção de ferramentas visuais de análise, sua arquitetura, características e utilização. Por fim, é descrito o *Pipeline* de visualização decorrente da junção entre o Arcabouço de visualização e a ferramenta FastMapDB.

O trabalho se encerra com uma breve análise da ciência de Visualização de Informações com base na literatura estudada, sendo traçado um cenário do estado da arte desta disciplina com sugestões de futuros trabalhos.

Abstract

In the present document are joined the collaborations of many works from the fields of Databases, Knowledge Discovery in Databases, Data Mining, and Computer-based Information Visualization, collaborations that, together, define the structure of the research theme and the work of the Masters Dissertation presented herein. This research topic is the Information Visualization discipline, and its relevant theory is reviewed and related to support the concluding activities, both theoretical and practical, reported in this work.

The referred work, anchored by the theoretical substance that was studied, makes several contributions to the science in investigation, the Information Visualization, presenting them through formalized proposals described across this text, and through practical results in the form of software enabled to the visual exploration of information. The presented ideas are based on the visual exhibition of numeric analysis, named basic statistics, frequency analysis (Frequency Plot), and according to a relevance analysis (Relevance Plot). There are also reported the contributions to the FastMapDB tool, a visual exploration tool built by the Grupo de Bases de Dados e Imagens do ICMC-USP, the performed enhancements are listed as achieved results in the text. Also, it is presented the Framework, as previewed in this work's original proposal, projected to allow the construction of visual analysis tools; besides its description are listed its architecture, characteristics and utilization. At last, it is described the visualization Pipeline that emerges from the joining of the visualization Framework and the FastMapDB tool.

The work ends with a brief analysis of the Information Visualization science based on the studied literature, it is delineated a scenario of the state of the art of this discipline along with suggestions for future work.

Capítulo 1 - Introdução

1.1 - Considerações Gerais

Uma quantidade sempre crescente de dados, oriundos das mais variadas áreas, têm sido gerados. O passo deste crescimento pode ser compreendido ao se afirmar que uma quantidade de dados da ordem dos bilhões de bytes foi gerada no curto espaço de tempo dedicado a este trecho de leitura. Assim, estima-se que ao decorrer de um ano o montante de informações criadas irá extrapolar inverossímeis exabytes (1 milhão de terabytes) de magnitude (Keim 2002). Esse universo de dados é armazenado em diversas modalidades de mídias, majoritariamente mídias de natureza digital passíveis de processamento em computadores. E, um melhor aproveitamento deste imenso volume de registros pode ser conseguido ao ampliar-se a compreensão da informação coletada, através da identificação inter-relacionamentos entre os elementos de dados. Desta forma, a busca por meios que possibilitem a descoberta de conhecimento inerente, e não aparente, em conjuntos de dados é o objeto de estudo do trabalho aqui desenvolvido.

No mundo contemporâneo, as informações são constantemente capturadas de forma automática devido à crescente participação dos computadores na sociedade, em todos os seus aspectos. Seja na realização de uma chamada telefônica ou na utilização de um cartão de crédito, temos computadores alimentados repetidamente por sistemas de monitoramento e sensores. Nos negócios, engenharia, ciência, medicina, nas instituições governamentais e comerciais, há um exército de máquinas prontas a anotar o que foi realizado, negociado, calculado, medido e transacionado. A atividade humana é informação, o máximo que se possa rastreá-la. No entanto, esta enxurrada de bytes é fonte de conhecimento ainda subexplorada e, conseqüentemente, decisões são tomadas a todo instante sem que todo conhecimento relevante seja levantado. Tais decisões podem ser não ótimas ou mesmo erradas (Keim, Ankerst et al. 1995).

A importância da busca de auxílio no aproveitamento das informações se torna ainda mais clara ao notarmos que o ritmo do armazenamento digital, em ordem de exabytes anuais,

resultará, no decorrer de três anos, em uma quantidade de informação superior a tudo que se criou em toda história prévia da humanidade (Keim 2002). Não com a mesma qualidade, note-se, mas com grande potencial. Potencial de mostrar características desconhecidas inerentes à natureza dos dados, potencial de proporcionar vantagens em um dado contexto, potencial de revelar, elucidar, concluir, potencial de informar e decidir. Mas, nada deste poder latente pode ser aproveitado sem as ferramentas adequadas.

1.2 - Motivação e Objetivos

Para que uma parcela maior das informações reunidas alcance a finalidade para a qual foram coletadas, o homem conta com o poder oferecido pelos sistemas de hardware computacionais, de natureza digital, adequados ao processamento maciço de informação. No entanto, para que esta facilidade possa ser aproveitada são necessários softwares capazes de promover a investigação dos dados armazenados. Porém, o desenvolvimento destes softwares, até então, ainda não é capaz de decifrar grandes quantidades de dados, pois carecerem de meios mais poderosos de investigação. Portanto, as ferramentas de exploração de dados a serem desenvolvidas devem buscar escalabilidade e poder investigativo, este último só podendo ser alcançado através de engenhosas *interfaces* de interação com o homem, pois se sabe que o processo de descoberta não pode ser totalmente automatizado (Keim, Ankerst et al. 1995) já que engloba inteligência e criatividade, características que o computador ainda não é capaz de simular. Ou seja, o homem ainda irá atuar decisivamente na utilização destes sistemas, que devem auxiliá-lo adequadamente (Schneiderman 1996).

Nesta perspectiva se encaixa a especialidade da ciência de computação denominada Knowledge Discovery in Databases (*Knowledge Discovery in DataBases* - KDD), um processo complexo que objetiva extrair conhecimento a partir de grandes volumes de dados. O KDD é um processo de investigação constituído por várias etapas: seleção, pré-processamento, transformação, Mineração de Dados (MD) e interpretação/avaliação (Fayyad, Piatetsky-Shapiro et al. 1996). Sua demanda vem impulsionando, principalmente, as pesquisas por novas técnicas de Mineração de Dados, que é o núcleo de todo processo.

Dentre as técnicas de MD que são objetos de pesquisa, há as técnicas de natureza visual que constituem a área de pesquisa denominada Visualização de Informações Auxiliada por

Computador. A essência da Visualização de Informações é possibilitar a introspecção de um conjunto de dados através de uma representação visual criada a partir da informação sendo investigada. Enquanto as técnicas de MD automatizadas não são capazes de evidenciar fatos interessantes como aglomerados e correlações, se estes não estiverem previstos pelos algoritmos que as definem, as técnicas visuais podem ressaltá-los naturalmente, sendo que seus limites são definidos pela habilidade dos analistas que as utilizam.

A utilização de meios visuais para exploração de dados é um meio poderosíssimo de se promover a descoberta de conhecimento. Isto se deve à combinação do enorme poder de processamento dos computadores atuais com a habilidade nata do cérebro humano em compreender e absorver a informação visual. O sistema de visão humano consegue, de maneira eficiente, destacar padrões e aspectos interessantes em cenas gráficas bem elaboradas derivadas de conjuntos de valores (Rao and Card 1994), ao passo que o computador é capaz de processar milhões de dados por segundo. Nesta abordagem, o homem e o computador amenizam mutuamente suas deficiências: a incapacidade de processar grandes volumes de dados e a incapacidade de analisar habilmente informações visuais, respectivamente. Juntos, portanto, compreendem um robusto mecanismo de investigação.

Na presente dissertação são apresentados os resultados do trabalho que culminou em um Arcabouço de visualização de dados que intenciona potencializar o processo de KDD, isto é, que o torne mais esclarecedor ao possibilitar a investigação mais eficaz dos conjuntos de dados. O trabalho apresentado combina uma série de técnicas de visualização altamente integradas. O resultado desta integração é a soma das principais vantagens de cada uma das abordagens ao mesmo tempo em que o efeito de suas fraquezas é atenuado por consequência das alternativas disponibilizadas. As técnicas do trabalho descrito se complementam e perfazem uma ferramenta altamente adequada para o efetivo aproveitamento dos dados ao possibilitar a geração de conhecimento.

1.3 - Apresentação e Organização do Trabalho

Este trabalho visa reunir amplas informações da área de banco de dados, permeando a especialização de Mineração de Dados orientada à Visualização de Informações, no contexto do KDD. Para tanto, cada um destes tópicos é abordado em diferentes níveis de detalhamento, sendo

que a Visualização de Informações tem especial atenção por se tratar do tema da proposta aqui desenvolvida.

O estudo teórico pormenorizado é o embasamento para a apresentação dos resultados do trabalho prático realizado. Este trabalho prático realizou-se na forma de diversas atividades de projeto, implementação, experimentação e validação que buscaram gerar novas contribuições à pesquisa da área de Visualização de Informações. Além disso, visou a integração de novas funcionalidades ao sistema em desenvolvimento pelo Grupo de Base de Dados e Imagens (GBdI) do Instituto de Ciências Matemáticas e de Computação (ICMC) da USP de São Carlos.

A organização desta dissertação é a seguinte:

O Capítulo 1 traz a introdução, motivação e a apresentação deste texto. O Capítulo 2 apresenta a base da disciplina de KDD, que possui como principal ramo a pesquisa por técnicas de Mineração de Dados. E, devido a esta profunda relação entre os termos KDD e MD, que muitas vezes são utilizados indistintamente, ambos os tópicos são tratados em um mesmo capítulo que procura elucidar os conceitos envolvidos e os principais progressos alcançados.

O Capítulo 3 trata do tema de Visualização de Informações. Neste capítulo há uma orientação do tema de Visualização no contexto de KDD e MD, ao mesmo tempo em que se busca clarificar as definições e conceitos da área. É realizada ainda, uma revisão da taxonomia proposta na literatura e uma exposição de técnicas ilustrativas da teoria estudada, bem como uma descrição de diversos subtópicos relevantes ao tema.

O Capítulo 4 apresenta os resultados alcançados. Estes resultados, o software, as técnicas e as contribuições, são expostos de forma descritiva abordando-se os detalhes de execução e a teoria utilizada como base. É realizada, ainda, uma discussão das implicações da pesquisa de Visualização de Informações no contexto do trabalho realizado.

Finalmente, no Capítulo 5, são feitas as considerações finais do trabalho através de uma síntese da teoria apresentada e dos resultados e contribuições alcançados, além disso é feita uma caracterização dos trabalhos futuros que possam vir a estender este projeto.

Capítulo 2 - Descoberta de Conhecimento em Bases de Dados

2.1 - Considerações Gerais

A disciplina denominada de Descoberta de Conhecimento em Bases de Dados (KDD) objetiva identificar nos dados, padrões, modelos ou estruturas válidas, novas, e potencialmente úteis que sejam sobretudo interpretáveis (Fayyad, Piatetsky-Shapiro et al. 1996). O KDD é composto por várias etapas, nas quais os dados são manipulados visando sua preparação para a principal atividade de extração de conhecimento, essa atividade é a Mineração de Dados, que na literatura muitas vezes é confundida com o próprio KDD, dada sua importância no processo completo. Portanto, o KDD (veja figura 1) é um procedimento complexo orientado à informação que objetiva gerar conhecimento tendo como fonte dados selecionados, processados e transformados para, finalmente, serem submetidos a técnicas específicas de Mineração de Dados. A interpretação dos resultados conseguidos é o produto final deste processo.

Figura 1 - As etapas que constituem o processo de KDD (extraído de (Fayyad, Piatetsky-Shapiro et al. 1996)).

2.2 - As Etapas do KDD

Os estágios do processo de Descoberta de Conhecimento em Bases de Dados compreendem atividades inerentemente iterativas, isto é, os resultados podem ser aprimorados através da utilização destes mesmos resultados para realimentar os próximos passos da iteração. Portanto, o analista de informações tem participação constante no processo, ele é quem indica quais são os melhores dados para se iniciar a análise, o grau de integridade em que devem se encontrar e em que formato poderão ser mais bem processados pelas ferramentas de Mineração de Dados. O analista, consciente de quais conhecimentos são valiosos, é quem dirige a MD e gera conclusões a partir do que lhe é apresentado na etapa final. Segundo (Fayyad, Piatetsky-Shapiro et al. 1996), as atividades iterativas e iterativas do KDD podem ser assim descritas:

- **Seleção:** os dados são escolhidos como um subconjunto de interesse, ou sumarizados em um subconjunto de amostragem. Os itens mais apropriados, segundo o analista, são colhidos e prossegue-se com a descoberta de conhecimento;
- **Pré-processamento:** são utilizadas ferramentas para preparação dos dados visando deixá-los em um formato mais apropriado para as próximas etapas. Aqui são tratadas distorções, ausência de dados ou, simplesmente, é realizada uma reorganização das informações;
- **Transformação:** os dados são processados e disponibilizados em uma forma diferente da original, mas ainda mantendo suas propriedades. O novo formato deve ser mais bem aproveitado para as etapas seguintes. Como exemplo, temos a discretização de dados, normalização e redução de dimensionalidade;
- **Mineração de Dados:** é extraído conhecimento com o auxílio de métodos computacionais capazes de revelar padrões, estruturas, tendências, etc. Vários métodos podem ser usados em função da natureza dos dados e das informações que se desejam alcançar, como identificação de aglomerados, geração de resumos e classificação;
- **Interpretação:** o resultado da mineração é submetido à apreciação do analista, que pode julgar necessário refazer o processo, alterando uma ou todas as etapas anteriores. Podem ser utilizados outros conjuntos de interesse, outras técnicas de pré-processamento/transformação ou ferramentas de MD adicionais.

A condução do KDD pode requerer a repetição dos processos até que se alcancem conclusões úteis a partir da massa de dados. Além disso, a participação do analista acontece durante todo o processo e é decisiva na aquisição de tais conclusões. Dessa forma, o procedimento é caracterizado como iterativo e interativo. A adequada realimentação do sistema, em seus vários estágios, é requisito para o sucesso da atividade de KDD. Para tanto, é necessário que o usuário compreenda o domínio dos dados e saiba quais objetivos almeja.

2.3 - Mineração de Dados

O conceito de Mineração de Dados pode ser definido como a descoberta de informações potencialmente úteis a partir de um conjunto de dados disponibilizado, de tal forma que tais informações não possam ser determinadas pela análise simples e direta. Por consequência, a atividade de MD não é trivial, sendo que a valia de sua utilização se deve ao fato de que seus resultados permitem uma melhor compreensão do conjunto de dados (Frawley, Piatetsky-Shapiro et al. 1991).

Segundo (Keim and Kriegel 1996), a idéia de MD pode ser formalmente definida como a busca por dois elementos:

- um subconjunto D' , pertencente a um conjunto de $D = \{d_1, d_2, \dots, d_n\}$;
- hipóteses $Hu(D', C)$ sobre D' , tais que o usuário as considere úteis no contexto de uma aplicação C .

A MD é um conjunto de técnicas e métodos que busca a identificação de relacionamentos e padrões existentes no conjunto de dados, auxiliando a tomada de decisões. Devido ao aumento extraordinário na aquisição e armazenagem de dados, e à motivação de se conseguir vantagens em qualquer atividade humana onde haja concorrência, o interesse em ferramentas de MD tem crescido muito nos últimos anos, o que se nota no aumento do número de publicações e de ferramentas de MD no decorrer do tempo. São trabalhos provindos principalmente das áreas de estatística, banco de dados, inteligência artificial, visualização, otimização e computação paralela.

A linha de MD originária da estatística é orientada a inferir padrões ou modelos a partir dos dados baseando-se em hipóteses. Em constraste, a MD, como é colocada na literatura atual, baseia-se em uma abordagem orientada a descoberta, onde não há necessariamente uma hipótese

prévia para um problema sob investigação. A MD não é uma finalidade, mas sim um meio através do qual se alcança um objetivo, seus benefícios provêm da aplicação dos resultados conseguidos sobre uma estratégia de negócios objetivando uma meta específica (Hirji 2001).

2.3.1 - Principais Abordagens da Mineração de Dados

A grande quantidade de algoritmos de MD já apresentados na literatura impede uma análise das técnicas de mineração focando-se na estrutura de seus princípios de implementação. Um tratamento mais adequado do tema é possível concentrando-se nos principais problemas abordados pelos algoritmos. Segundo (Chen, Han et al. 1996), os principais métodos que compõem a aplicação de MD são:

- **Regras de associação:** busca descobrir regras para inferência dos dados da seguinte forma: *se $A_1 \wedge A_2 \wedge \dots \wedge A_m$ então $B_1 \wedge B_2 \wedge \dots \wedge B_n$* . Onde A_i ($i \in \{1, \dots, m\}$) e B_i ($i \in \{1, \dots, n\}$) são conjuntos de valores de atributos do conjunto de dados relevantes da base de dados. Por exemplo, pode-se inferir em uma base de dados de um supermercado que quando os clientes compram cerveja, geralmente compram carne de churrasco;
- **Generalização e sumarização:** como o próprio nome diz, procura gerar uma caracterização, uma visão geral de um conjunto de dados fornecido. Por exemplo, a partir do banco de dados do supermercado do exemplo anterior, pode-se caracterizar que os clientes que compram cerveja e carne de churrasco são casados, com mais de 30 anos e pertencem a uma determinada faixa salarial;
- **Classificação:** procura classificar os dados de um conjunto baseando-se nos valores de alguns atributos. Como exemplo, os compradores do supermercado poderiam ser separados em classes de acordo com suas preferências por marca de cerveja e tipos de carne. O supermercado poderia, a partir destes resultados, criar promoções específicas para estes consumidores;
- **Identificação de aglomerados:** também chamado de segmentação, procura particionar os elementos de dados em subconjuntos cujos elementos possuem uma certa similaridade, de forma que os dados com propriedades semelhantes são considerados homogêneos. Através deste tratamento dos dados, os clientes de um supermercado

podem ser divididos em um número finito de categorias, cada uma das quais com tendências de compra e periodicidade;

▪ **Busca de padrões em dados temporais:** procura identificar padrões que apenas se manifestam ao longo do tempo. Os resultados são utilizados para previsão de risco, identificação das causas de determinados fenômenos e tendências associadas a padrões. Alguns exemplos de resultados esperados através da utilização destas técnicas são a identificação de produtos com variação de preços similares, ações com valorizações semelhantes e empresas com crescimentos semelhantes.

Observando-se as diferentes finalidades dos algoritmos de MD, fica clara a abrangência de suas aplicações e como seus resultados podem ajudar em diversos domínios de aplicação. A utilização dos resultados obtidos promove vantagens nos negócios, conclusões em estudos científicos, prevenção de riscos, previsão de fenômenos, etc. Mas, apesar da polivalência das técnicas de MD, seus princípios de aplicação ainda requerem aperfeiçoamentos para que se tornem mais acessíveis e eficazes. Isto se deve ao fato de que a maioria das ferramentas de MD possui pouca ou quase nenhuma participação do usuário durante o processo de descoberta, fazendo com que a Mineração de Dados se torne, muitas vezes, um processo maçante e dispendioso. O analista é obrigado a repetir os processos várias vezes com diferentes conjuntos de dados e parâmetros iniciais que não podem ser alterados no decorrer da execução. A sua utilização prende-se a ciclos de tentativa e erro até que resultados possam ser aproveitados (Ganesh, Han et al. 1996). A seguir será revista uma proposta de metodologia para aplicação de MD em meios empresariais visando otimizar o processo de descoberta de conhecimento.

2.3.2 - Uma Proposta de Metodologia para Aplicação da Mineração de Dados

Em (Cabena, Hadjinian et al. 1998) apud (Hirji 2001) é apresentada uma metodologia de aplicação através de um método composto de cinco estágios que descrevem como realizar a MD. Em (Hirji 2001) é apresentado um caso de teste seguindo esta metodologia; neste trabalho são descritas todas as etapas, os problemas e soluções encontradas no contexto de uma empresa em busca de maior competitividade através do aproveitamento de dados previamente

armazenados. As etapas, descritas com detalhes, constituem uma transcrição do procedimento de KDD, como descrito na seção 2.1, para o ambiente corporativo. Os estágios previstos são:

- **Determinação dos objetivos de negócios:** nesta etapa são identificados os objetivos a serem alcançados, os dados a serem utilizados e os propósitos do conhecimento a ser extraído;
- **Preparação dos dados:** consiste das tarefas de seleção dos dados, pré-processamento e transformação. Os dados são filtrados, "limpos", organizados e/ou alterados objetivando resultados ótimos de performance e geração de resultados na próxima etapa. Geralmente, é a etapa que consome maior esforço em todo processo;
- **Mineração dos dados:** abrange a seleção e a execução dos algoritmos de MD. Nesta etapa o especialista de MD desempenha um papel crucial para o sucesso das etapas seguintes;
- **Análise dos resultados:** são avaliadas as saídas geradas pelo processo até este ponto, a qualidade dos resultados é medida baseando-se na utilidade e originalidade das informações;
- **Assimilação de conhecimento:** a etapa final procura adaptar o conhecimento gerado com as metas inicialmente traçadas. Novas metas também são determinadas a partir da identificação de resultados potencialmente úteis.

Neste estudo de caso, originário da aplicação da MD em um ambiente corporativo, é ilustrada a complexidade da aplicação do processo de geração de conhecimento. Nota-se a necessidade de um minucioso planejamento antes do início da execução para que todas as demais etapas sejam guiadas para metas previamente estipuladas. A preparação dos dados, como descrito em (Hirji 2001) é uma etapa crítica da atividade de MD, pois se os dados não estiverem armazenados de forma sistemática, a continuação do trabalho pode ser impedida ou muito atrasada. No trabalho analisado, a mineração mostrou-se como a mais importante etapa para obtenção dos objetivos iniciais, ao mesmo tempo em que também pôde ser considerada a mais complexa do ponto de vista técnico, já que a escolha dos algoritmos é que determina o sucesso do processo. A análise dos resultados é outra etapa cuja preparação técnica tem importância, já que a interpretação dos resultados nem sempre é compreendida pelos clientes do trabalho de mineração. Além disso, o conhecimento adquirido deve ser não trivial enaltecendo a qualidade do

processo com a apresentação de valiosas conclusões não esperadas, evitando que o cliente reaja com um "isso nós já sabemos". Na finalização, os resultados aproveitados são confrontados com as metas iniciais e um plano é traçado visando utilizar a compreensão adquirida, ou refinar o que foi alcançado através de uma nova iteração.

Embora a metodologia proposta não possa ser considerada uma abordagem absoluta para a execução da MD, determinando o encerramento das pesquisas por outros modelos, sua contribuição é capaz de elucidar a complexidade do problema. Ao mesmo tempo, o estudo propõe uma solução resultante do tratamento analítico do tema resultando em uma seqüência de procedimentos enumerados que, se cumpridos, podem culminar no sucesso do processo de MD.

2.3.3 - Arquiteturas de Sistemas de Mineração de Dados

Os sistemas de Mineração de Dados são classificados em função do grau de integração que possuem com um sistema de banco de dados ou de um *data warehouse*. Eles podem determinar sistemas independentes, sem qualquer acoplamento com o banco de dados/*data warehouse*, ou podem ser sistemas integrados (Han and Kamber 2001). Se há integração, o grau de acoplamento varia de fracamente acoplado para fortemente acoplado.

Sistemas independentes, sem acoplamento algum, não se beneficiam das funções de análise, tratamento e manipulação de dados oferecidas pelos softwares de armazenamento em massa. Os dados são acessados em uma fonte externa ao banco de dados/*data warehouse*, como por exemplo, um simples arquivo texto ou binário, em seguida são processados e os resultados armazenados em um terceiro arquivo. Além disso, sistemas independentes de MD não se favorecem das propriedades dos softwares de armazenamento em massa, onde os dados permanecem organizados, integrados, indexados, limpos e íntegros. Nesta arquitetura, também não são aproveitados mecanismos já consolidados de acesso e armazenagem fundamentados em algoritmos e estruturas de dados amplamente testados.

Em (Lutu 2002) são descritos os graus de acoplamento. O acoplamento fraco indica que apenas algumas das facilidades dos bancos de dados/*data warehouses* são utilizadas pelo software de MD; os acoplamentos semifortes estendem este conjunto de funcionalidades e os acoplamentos fortes podem ser considerados unidades funcionais de um software que agora pode ser denominado Sistema de Informação.

No acoplamento fraco, algumas facilidades dos bancos de dados/*data warehouses* como carregamento de dados e armazenagem dos resultados no próprio sistema de dados já são utilizadas. Em um nível mais alto de acoplamento, o semiforte, o acoplamento vai além do aproveitamento do mecanismo de recuperação/armazenagem de dados. Neste grau de integração, a acoplagem é estendida ao fazer uso de funções como ordenação, indexação, agregação, análise de histogramas e pré-computação de estatística básica, soma, contagem, máximo, mínimo, média e desvio-padrão.

No acoplamento forte a MD faz uso das capacidades de consulta do banco de dados para efetuar as análises que resultarão nas conclusões do processo de mineração. O grau de integração é tão forte que a ferramenta de MD pode ser considerada uma funcionalidade a mais do banco de dados/*data warehouse* e não mais um software a parte que deve ser manipulado pelo analista de informações. Como uma capacidade extra de análise, que faz uso de todo o potencial do software de armazenagem em massa, a MD passa a ser uma comodidade computacional capaz de otimizar a armazenagem dos dados, intensificar sua utilização inteligente, aperfeiçoar sua organização, etc. A união do ferramental de análise com o poder do software de dados compreendem um verdadeiro Sistema de Informação.

2.4 - Implicações sobre um Sistema de Informação

Um Sistema de Informação pode ser definido como um arranjo de fatores computacionais especialmente arquitetados para construir, operar, gerenciar, exibir e reportar informações providas de uma base de dados. A sua característica primordial de auxiliar a análise e recuperação de informações em massa determina que algumas premissas orientem o desenvolvimento e o planejamento de sistemas desta natureza. Em (Chen, Han et al. 1996) são apontadas algumas características e desafios para que se alcance um Sistema de Informação:

- **Manipulação de diferentes tipos de dados:** dada a abrangência e magnitude dos SIs, espera-se que tais sistemas sejam capazes de manipular uma grande variedade de dados, provindos dos mais diversos domínios. Por manipulação, aqui neste contexto, entende-se que um efetivo processo de MD possa ser realizado neste cenário heterogêneo. No entanto, não se pode desejar que um SI seja capaz de manipular, em uma única implementação, todos os tipos de dados. Informações de natureza multimídia,

transacionais, hipertexto, estruturados, etc; dados originários tanto de bancos de dados relacionais, quanto de transacionais, e dados espaciais, podem determinar a necessidade de um SI com propriedades específicas;

- **Eficiência e escalabilidade de algoritmos de MD:** espera-se que os procedimentos de Mineração de Dados ocorram dentro de um período de tempo aceitável e previsível, ao mesmo tempo em que sejam escaláveis;
- **Usabilidade, precisão e expressividade dos resultados da MD:** a atividade de KDD deve ser capaz de meticulosamente descrever os dados armazenados, provendo medidas de precisão levantadas durante o próprio procedimento de MD. Dados ruidosos e/ou excepcionais não devem influenciar na qualidade das conclusões geradas;
- **Apresentação dos resultados em diferentes formatos:** o sistema deve apresentar os resultados alcançados pelo procedimento de KDD em diversos formatos, para que as conclusões possam ser apreciadas a partir de diferentes pontos de vista e por analistas com diferentes graus de preparação. Ferramentas gráficas e diferentes abordagens para expressar os resultados são necessárias;
- **Mineração interativa em múltiplos níveis de abstração:** como é difícil prever o que se pode descobrir a partir de um banco de dados, é necessário que *interfaces* bem trabalhadas sejam capazes de fornecer subsídios para que uma exploração eficiente e reveladora decorra sobre um ambiente de alto-nível de abstração. A interação deve ser observada de maneira a permitir que o usuário refine os objetivos de exploração, altere o foco dos dados, aprofunde-se no processo de MD e possa apreciar os dados e os resultados da mineração por diferentes ângulos e níveis de abstração;
- **Mineração a partir de diferentes fontes de dados:** o Sistema de Informações deve aproveitar a possibilidade de utilização de diversas e heterogêneas fontes de dados distribuídas e acessadas via infra-estrutura de rede. Ao mesmo tempo em que tal potencial não pode ser desprezado, sua utilização implica em um aumento de complexidade decorrente das diferentes apresentações dos dados, da localização esparsa das fontes e do imenso volume de informação acessada;
- **Proteção da privacidade e segurança dos dados:** os aspectos observados até aqui devem levar em consideração a segurança e privacidade do sistema, visto que a agregação destes fatores levanta vários aspectos potencialmente exploráveis do ponto

de vista de invasão de sistemas. Portanto, deve ser observado o constante zelo pela integridade e acessibilidade das informações armazenadas.

A exposição acima, embora de grande relevância, tem caráter simplista na discussão que um sistema destas proporções viria a ter. Um grande número de pormenores pode ser apontado no planejamento de um SI contemplando todas as características acima levantadas; muitas destas características apresentam complexidade capaz de compor verdadeiras linhas de pesquisa para tratar as implicações de sua execução. O último item observado, por exemplo, é apenas um dos conflitos que podem surgir quando se caminha para a satisfação plena de todos os preceitos descritos acima. Outro exemplo das dimensões que uma discussão sobre o tema pode levantar está na complexidade prevista nas características desejadas nos sistemas de SI, que supõem a manipulação de diferentes tipos de dados, provindos de diferentes fontes, ao mesmo tempo em que se promove uma mineração interativa por diferentes níveis de abstração. Portanto, o valor das informações apresentadas nesta seção está na elucidação das faculdades que um SI deve apresentar, o que promove um panorama do sistema ideal que se deseja; mas, ainda que pertinente ao estudo aqui proposto, a discussão aprofundada do tema não está no escopo do trabalho. Um debate mais aprofundado pode ser encontrado em (Clifton and Marks 1996).

2.5 - Visualização, Mineração de Dados e Mineração Visual de Dados

A Visualização de Informações, como descrita na seção 1.2, é um recurso com propriedades capazes de auxiliar fortemente o processo de descoberta de conhecimento. A visualização baseia-se na capacidade humana de interpretar informações visuais e, através dela, o homem é capaz de perceber com rapidez relacionamentos, padrões, tendências e exceções em uma representação visual de um dado conjunto de informações. A Mineração de Dados, como visto até aqui, também é uma ferramenta com grande potencial revelador. No entanto, possui um princípio de funcionamento diferente; baseado principalmente em análises estatísticas, a garimpagem dos dados é realizada de forma automatizada com participação reduzida do usuário durante a tarefa de mineração propriamente dita. Além disso, a automatização do processo, em detrimento da interação do usuário, pode gerar a necessidade de re-execução de todo o processo de MD para cada tentativa onde um novo conjunto de parâmetros será testado.

Desta forma, ao passo que a Visualização de Informações não é capaz de substituir as técnicas convencionais de MD, ao mesmo tempo oferece possibilidades únicas que não podem ser desprezadas. A utilização destas duas frentes de análise pode potencializar enormemente a exploração de informações, no entanto, a utilização intercalada, ao invés de integrada, das técnicas acaba por penalizar os dois procedimentos com as deficiências e limitações de cada um (Wong 1999). Segundo (Foster and Gee 2002), o processo de KDD não é uma caixa preta baseada em alguns algoritmos de busca capazes de retornar informações a respeito dos dados, mas sim um processo interativo que envolve o ser humano. Portanto, a utilização conjunta do KDD e da Visualização de Informações tem sido objeto de estudo em diversos trabalhos, como em (Hinneburg, Keim et al. 1999), que afirma que técnicas de visualização são capazes de melhorar consideravelmente a eficiência dos processos de identificação de aglomerados via MD. Neste cenário, surge a proposta de união destas duas frentes objetivando melhores resultados no processo de KDD; denominada Mineração Visual de Dados (MVD), esta ciência procura unir virtudes de duas áreas de pesquisa distintas, mas que possuem objetivos comuns e grande potencial de integração.

A Mineração Visual de Dados pode ser definida (Ganesh, Han et al. 1996) como a utilização de técnicas de visualização para que o usuário explorador das informações possa decidir mais facilmente quais dados de entrada escolher, compreender adequadamente os resultados e, além disso, avaliar, monitorar e guiar o processo de mineração. Nesta mesma linha, (Wong 1999) defende que certos passos matemáticos de um procedimento analítico podem ser substituídos por decisões humanas baseadas em visualizações dos dados, permitindo assim que este mesmo procedimento analítico possa analisar uma extensão mais ampla da informação. Ainda, segundo este mesmo autor, permitir que a visualização humana participe de um processo analítico com tomada de decisão ainda é um grande desafio. Em (Keim 2002) é lembrado que a grande vantagem da MVD é o fato de que o usuário está diretamente envolvido no processo de Mineração de Dados. Em (Ankerst, Ester et al. 2000) e (Ankerst, Elsen et al. 1999) são apresentadas as bases e os resultados de um trabalho de integração de técnicas visuais com algoritmos de classificação baseados em árvores de decisão.

Especificamente, o tópico do trabalho aqui desenvolvido é a Visualização de Informações. No entanto, pelo fato de que as técnicas de Visualização de Informações podem ser agregadas aos sistemas de KDD, seja como uma alternativa durante a etapa de mineração, ou

como um meio de potencializar a abordagem analítica das ferramentas de MD, então as técnicas de visualização devem ser estudadas em conjunto com o tópico de descoberta de conhecimento ao mesmo tempo em que são confrontadas com a MD convencional, assim como foi realizado. Portanto, o tratamento conjunto das áreas de pesquisa mencionadas conduz a uma investigação completa da ciência de Visualização de Informações, especialização esta que tem como motivo, comum e primordial, a investigação de grandes volumes de dados para geração de conhecimento.

2.6 - Considerações Finais

Neste capítulo o tema de KDD foi conceituado e analisado através das etapas que o constituem, sendo que a principal destas etapas, muitas vezes confundida com o próprio KDD é a Mineração de Dados. A Mineração de Dados em si constitui uma área de pesquisa ampla e que tem motivado pesquisadores de diversas modalidades, da estatística à inteligência artificial, da computação gráfica à interação homem-computador.

Devido à enorme abrangência e importância da MD, este tema também foi analisado através da apresentação de suas diversas abordagens, que são a base da maioria dos sistemas de KDD atuais. Uma metodologia de aplicação de MD também foi exposta como ilustração da complexidade e objetivos da utilização da Mineração de Dados. Por fim, as possibilidades de arquitetura de sistemas de MD foram revistas no contexto de bancos de dados, sendo descrito, ainda, como tais arquiteturas podem chegar a compor verdadeiros Sistemas de Informação: ambientes onde a coleta, o acesso e a exploração de dados são o principal objetivo.

Concluindo o capítulo foi feita uma confrontação do conceito de MD, que acabara de ser apresentado, com o conceito de Visualização de Informações que será apresentado em seguida. Esta comparação foi realizada em conjunto com o conceito de Mineração Visual de Dados para que uma compreensão mais completa dos temas pudesse ser alcançada.

Capítulo 3 - Visualização de Informações

3.1 - Considerações Gerais

A Visualização de Informações é a modalidade de Mineração de Dados que proporciona compreensão e análise da informação através de representações visuais construídas a partir dos próprios dados sendo investigados. As técnicas empregadas são capazes de desvendar enormes quantidades de dados com muita rapidez, propiciando suporte a analistas de informação na tarefa de alcançar uma introspecção mais profunda nos conjuntos de dados.

As técnicas de Visualização de Informações são propostas objetivando, principalmente, a investigação de conjuntos de dados de alta dimensionalidade. Segundo (Oliveira and Levkowitz 2002), o limite conceitual entre baixa e alta dimensionalidade está em torno de 34 atributos. Porém, dependendo da visão de cada autor, esse limite varia entre 5 a 10 (Beyer, Godstein et al. 1999) (Berchtold, Böhm et al. 1997) (Berchtold, Böhm et al. 1998) para até mais de 100 (Böhm and Kriegel 2000). Ainda segundo (Oliveira and Levkowitz 2002), ressaltando-se a capacidade de percepção humana, não há diferença inteligível entre um conjunto de dados com 5 dimensões e outro com 50 dimensões, já que ambos estão além da habilidade humana de compreensão baseada na analogia geométrica, restrita a 4 dimensões.

Os objetivos da Visualização de Informações, segundo (Keim 1997), dividem-se em função de três atividades de análise:

- **Análise exploratória:** sem nenhuma hipótese a respeito dos dados, o processo segue a esmo interativamente procurando por estruturas, tendências, etc;
- **Análise confirmativa:** com uma hipótese já formulada, prossegue-se através de um caminho cujo objetivo já é conhecido. A hipótese poderá ser confirmada ou rejeitada;
- **Apresentação:** fatos conhecidos *a priori* são apresentados com auxílio da ferramenta de visualização que provê um mecanismo eficiente de exibição.

3.2 - Visualização de Informações x Visualização Científica

Inicialmente, a utilização de meios visuais para análise de informações se beneficiava da representação geométrica/espacial inerente aos dados. Como, por exemplo, informações de até três dimensões que determinam curvas no espaço cartesiano, e informações de origem geográfica, cuja representação pode assemelhar-se ao espaço físico onde ocorreu a coleta de dados. No entanto, a maior parte das informações geradas pelos sistemas não possuem estas características, pois são inerentemente multidimensionais e complexos, e, geralmente, não possuem dimensões óbvias (Lux 1998). Portanto tais dados não podem ser descritos geometricamente, constituindo informações denominadas abstratas.

Estas duas classes de dados, geométricos e abstratos, são ambas de interesse da Visualização de Informações Auxiliada por Computador, ou simplesmente Infovis (Munzner 2002), mas em ramos diferentes desta ciência. Os dados geométricos são tratados pela Visualização Científica e os abstratos pela Visualização de Informações. A primeira utiliza uma representação espacial que está implícita nos dados, enquanto que a segunda procura definir um mapeamento gráfico dos dados que não são inerentemente espaciais.

É interessante observar que os nomes das subáreas da Infovis nasceram em decorrência de confusões históricas, como observa (Munzner 2002), e suas denominações incorporam conotações infortúnias quando justapostas: Visualização Científica não deixa de ser informativa, e Visualização de Informações não deixa de ser científica.

Um exemplo de Visualização Científica é o fluxo de ar sobre as asas de um avião, informação no formato de vetores 3D amostrados periodicamente em um experimento aeronáutico. Na Visualização Científica, a representação deste fenômeno seria, provavelmente, através de flechas indicando a direção do fluxo de ar posicionadas em relação às asas do avião. Com as cores dos vetores possivelmente indicando a velocidade do fluxo. Generalizando, a Visualização Científica trabalha visando potencializar o sistema sensorial humano, seus resultados buscam exibir fenômenos que são muito rápidos ou muito lentos para que o olho os perceba, ou mesmo estruturas muito maiores ou menores do que a escala da visão humana, como estruturas protêicas.

Já a Visualização de Informações pode ser exemplificada através de uma base de dados de clientes de um supermercado, com nome, faixa salarial, ano de nascimento e sexo. Este

conjunto de dados pode ser considerado abstrato já que não possui representação espacial implícita, ao invés disso, necessita de um mapeamento para uma representação gráfica, adequada à noção espacial do ser humano e capaz de revelar dados interessantes para um determinado propósito. De maneira geral (Duke 2001), a Visualização de Informações engloba dados que não possuem uma representação natural para um espaço R^n onde, geralmente, $n \leq 3$.

É interessante notar que qualquer dado utilizado na Visualização Científica pode ser utilizado com sucesso na Visualização de Informações, o inverso não é sempre verdade. Pode-se dizer, portanto, que Visualização de Informações tem um caráter mais genérico, enquanto que a Visualização Científica é mais específica, podendo muitas vezes ser encarada como uma reprodução de algum fenômeno cujos dados foram coletados.

Uma nota relevante, que questiona a separação da Infovis em duas subáreas, é o fato de que estes dois tratamentos, previstos para introspecção visual sobre dados, possuem uma série de propriedades comuns: têm o mesmo objetivo global, utilizam os mesmos recursos computacionais, podem ser igualmente agregados a um Sistema de Informação, são baseados no mesmo princípio de representar dados graficamente, são orientados ao usuário e utilizam interatividade para alavancar seus resultados. Além disso, em diversos casos, as diferenças entre as subáreas são tão tênues, no que tange à natureza dos dados, que chega a ser possível sua utilização intercalada e mesmo complementar. É o que se observa em (Duke 2001) que, tomando proveito da natureza comum das duas abordagens, propõe uma metodologia de desenvolvimento visando integrar ambas as técnicas.

O presente trabalho versa sobre a ciência de Infovis, mais especificamente na subárea de Visualização de Informações como já definida, tratando de dados com maior dimensionalidade.

3.3 - Técnicas de Pré-processamento dos Dados

As técnicas de pré-processamento têm como função melhorar a visualização dos dados aplicando a estes processamentos tais que o conjunto de dados possa ser particionado, selecionado, sintetizado, diminuído ou colocado em formato mais adequado para apresentação, mas que preserve as características do conjunto original. O pré-processamento pode ser realizado antes ou mesmo durante a visualização, através de técnicas de interação que permitam ao usuário

manipular as informações para que sejam exibidas de maneira mais esclarecedora. Segundo (Keim 1997), as técnicas de pré-processamento podem ser divididas em:

- **Técnicas de redução de dimensionalidade:** um dos grandes problemas da MD surge quando o número de dimensões é muito alto, fazendo com que o processamento fique sobrecarregado ou que técnicas de visualização tornem-se impraticáveis, pois perderiam a inteligibilidade. Para amenizar esta adversidade, as técnicas de redução de dimensionalidade transformam os dados em um formato mais simples, mas que busca preservar as propriedades originais. Em uma forma mais sucinta, as informações são mais bem apresentadas, compreendidas e processadas. Como exemplo temos os trabalhos (Sammon 1969) e (Faloutsos and Lin 1995);
- **Técnicas de seleção de dados:** determinam a utilização de apenas uma parte da base de dados na visualização, por amostragem de dados representativos, ou por seleção através de uma consulta a uma base de dados;
- **Técnicas de segmentação:** particionam a base de dados segundo um ou mais atributos, as partições resultantes podem ser analisadas separadamente;
- **Técnicas de agregação:** neste tratamento, os dados são sumarizados por funções de agregação como soma, contagem, mínimo, máximo, etc, aplicadas sobre os valores dos atributos. Estas técnicas são as mais antigas e conhecidas, e servem de subsídio para visualizações clássicas como histogramas, gráfico de barras, entre outras.

Apesar da denominação de pré-processamento, não há uma convenção sobre quando estas técnicas são aplicadas. Como veremos neste trabalho, os dados podem ser selecionados durante a exibição da visualização, agregações podem ser requisitadas via interação do usuário sobre a cena já criada, ou os dados podem ser selecionados através de uma técnica de visualização para serem mais bem explorados através de uma segunda técnica, determinando um *pipeline* de visualização. Resumindo, as técnicas atuam sobre o conjunto de dados inicial que pode ser transformado em tamanho, forma ou em seus valores; são operações que mantêm as características originais da base e que podem ser operadas em diversos momentos da visualização.

3.4 - Interação

A complexidade dos dados tratados na Visualização de Informações, desprovidos de representação geométrica e até mesmo de semântica, exige paradigmas mais poderosos para a investigação visual auxiliada por computador, paradigmas não mais restritos à utilização de simples representações gráficas estáticas. E, como se pode afirmar que uma imagem vale por mil palavras, então uma imagem interativa vale por muito mais. É sobre este princípio que a ciência da Visualização de Informações define as diretrizes que a guiam no objetivo de dominar a análise de dados complexos: a Visualização de Informações deve ser interativa. Neste contexto, uma definição mais completa pode ser formulada: a Visualização de Informações baseia-se na criação e no projeto de representações gráficas interativas da informação combinando princípios das disciplinas de projeto gráfico, ciência cognitiva e computação gráfica interativa (Chi 2002).

A exploração de bases de dados é inerentemente interativa (Ho and Nguyen 2001), concentrando-se em refinar hipóteses baseadas em resultados decorrentes da interação com os dados através de ferramentas de análise, consultas a bancos de dados, visualização, etc. A interação, que pode ser encarada como o diálogo entre o usuário e a máquina, é capaz de potencializar enormemente o poder elucidativo de uma dada técnica de visualização. Interagindo dinamicamente, o usuário pode alterar a visualização de forma que suas metas de exploração possam ser alcançadas. Meios adequados de interação permitem ao usuário criar diversos arranjos da estrutura da base de dados sendo explorada, comparar suas dimensões e gerar conhecimento a partir da análise das projeções geradas em cada passo do processo.

3.4.1 - Técnicas de Interação

A classificação do conhecimento gerado em qualquer área de pesquisa é importantíssima, pois proporciona uma organização das idéias, estruturação de metodologias, posicionamento de novas propostas, delineamento de progressos conseguidos e futuros. No entanto, padrões de classificação em uma dada ciência demoram a se consolidar, seja devido à imaturidade das descobertas, devido a novas descobertas, a divergências entre pesquisadores, e até mesmo competição intelectual, entre outros interesses. Em muitos casos, padrões absolutos nunca são alcançados, determinando que um certo ramo de pesquisas fique fadado a conviver com diversas

denominações de um mesmo conceito. O mesmo vale para a Infovis, uma ciência relativamente jovem na qual novas contribuições são incessantemente propostas anualmente.

Desta forma, nesta seção procura-se delinear a natureza das técnicas de interação do estado da arte da disciplina por duas visões. Na primeira, publicada em (Keim 2002), são descritas classes bem definidas nas quais as técnicas devem se encaixar; na segunda, em (Grinstein and Ward 2002), não há uma classificação explícita, mas sim uma discussão dos fatores que podem ser utilizados para agrupar os dados segundo propriedades comuns. Portanto, visando uma revisão mais abrangente do tema de classificação, nesta seção e na seção 3.5, que aborda classificação de técnicas de visualização, serão apresentados os pontos de vista de ambos autores.

Segundo (Keim 2002), as técnicas de interação se dividem em projeção interativa, filtragem interativa, zum interativo, distorção interativa e *Link & Brush*. Tais técnicas são resumidas a seguir:

- **Projeção interativa:** possibilita a redefinição dinâmica de projeções geradas a partir de um conjunto multidimensional. A geração de diferentes projeções, pela intervenção do usuário, deve promover a elucidação gradativa da base de dados. Como exemplo temos (Wegman and Luo 1997), com o sistema GrandTour, que procura exibir todas as projeções bidimensionais de um conjunto multidimensional como uma série de *scatter plots*;
- **Filtragem interativa:** é o recurso mais utilizado em técnicas de visualização de todas as naturezas. Trata-se da possibilidade de realizar consultas visuais sobre os dados em análise; assim, o usuário é capaz de focar em porções dos dados que julgar mais interessantes, gerando diferentes visualizações que podem ser comparadas e utilizadas para elucidar a relação entre conjuntos de dados selecionados sobre diferentes consultas. Exemplos que fazem uso desta técnica podem ser observados em (Rao and Card 1994) e (Rundensteiner, Ward et al. 2002), e em (Martin and Ward 1995) onde a idéia é analisada minuciosamente;
- **Zum interativo:** é outra maneira de focar em diferentes porções da base de dados, mas neste caso diferentes visões da distribuição dos dados são conseguidas de acordo com o grau de zum. Isto é possível comprimindo-se os elementos de visualização (*pixels*, linhas, ícones ou qualquer marca gráfica) e expandindo (detalhando) apenas aqueles onde houver interesse. Maiores níveis de zum correspondem a mais detalhes, segundo

a manipulação do usuário. Um exemplo de implementação deste modo de interação é apresentado em (Bier, Stone et al. 1993) e (Rao and Card 1994);

- **Distorção interativa:** através desta técnica é possível ter-se uma visão global da presença e distribuição dos dados. O recurso utilizado é a deformação dos elementos visuais de maneira que suas propriedades espaciais ainda permaneçam claras (ver figura 2). O usuário fica habilitado a navegar através da estrutura espacial do conjunto de dados e visualizar os detalhes que julgar mais interessantes. Um estudo profundo das técnicas de distorção pode ser encontrado em (Leung and Apperley 1994) e exemplos de implementação de técnicas desta natureza podem ser conseguidas em (Mackinlay, Robertson et al. 1991) com a *Perspective Wall*, em (Sarkar and Brown 1994) com a *Fish-eye View*, e em (Walter and Britter 2002) com o plano hiperbólico;

Figura 2 - Modelo mecânico (a) que caracteriza as técnicas *Perspective Wall* e *Bifocal Displays*. Em (b) é apresentada a aparência do espaço de dados transformado pela técnica ilustrada em (a). Extraído de (Leung and Apperley 1994).

- **Link & Brush:** ou co-plots (Wegman and Luo 1997), é um método extremamente importante relacionado às técnicas de interação acima descritas. Como as diversas técnicas de visualização possuem pontos positivos (aptidões) e fraquezas distintas, peculiares aos diferentes esquemas de apresentação dos elementos visuais, é interessante que um mesmo conjunto de dados possa ser analisado por diferentes abordagens, tanto em totalidade quanto em regiões de interesse definidas interativamente. Desta forma, tendo-se um conjunto de dados como fonte para diversas técnicas de visualização apresentadas simultaneamente, seu princípio é propagar as ações do usuário para todas as representações visuais do conjunto de dados que está sendo analisado (Ward 1997).

De acordo com a segunda abordagem de classificação das técnicas de interação (Grinstein and Ward 2002), o usuário e os dados em uma cena de visualização podem fazer uso da **interação através de recursos de navegação**, isto é, alternando parâmetros gráficos que lhe permitam ver a imagem por diferentes ângulos buscando um quadro mais revelador. O usuário pode também utilizar a **interação por amostragem dos dados** para reduzir as proporções do processo de análise que viria a ser realizado sobre um conjunto menor de informações. Há também a **interação direta**, através da qual é possível fazer consultas (*queries*) para fins específicos que surjam durante o processo de análise. E, por fim, é apontada a **interação associativa** que permite o acesso relacionado dos dados em diferentes técnicas de visualização. Ainda em (Grinstein and Ward 2002), é levantada a possibilidade de **interação sobre o sistema que suporta a visualização**, isto é, sobre a rede que o assiste, os arquivos fonte e os que venham a ser gerados, sobre parâmetros de funcionamento como os atributos a serem considerados na visualização, exibição de dados complementares, etc.

Fica claro nos parágrafos acima que os pontos de vista dos trabalhos estudados, (Keim 2002) e (Grinstein and Ward 2002), têm muitos pontos em comum, não sendo difícil traçar uma relação entre eles. Mas, mesmo assim, ambos propõem contribuições inéditas para o objetivo de categorizar as técnicas interativas. O primeiro consegue definir um pequeno conjunto de classes dentro das quais praticamente todos os trabalhos de interação já descritos podem ser encaixados. O segundo propõe caracterizações não tão claras, mas com grande potencial de utilização, sendo que sua maior contribuição foi chamar a atenção para outros aspectos de interação que não a manipulação da cena propriamente dita.

3.4.2 - Um Modelo de Interação

Em (Keim, Lee et al. 1995) é feita uma análise interessante do processo de interação usuário-computador no contexto dos sistemas de mineração visual de dados. Inicialmente, a proposta isola os componentes participantes do processo e faz uma análise da inter-relação entre eles buscando um resultado altamente integrado para uma melhor usabilidade.

Os componentes do sistema descrito são:

- **Gerenciamento de dados:** é o sistema gerenciador de banco de dados propriamente dito, capaz de gerenciar automaticamente a armazenagem em massa, prover a

recuperação eficiente dos dados através de uma linguagem de manipulação, e efetuar controle de concorrência mantendo a integridade da informação;

- **Análise dos dados:** são ferramentas de análise automatizada capazes de realizar levantamentos estatísticos sobre os dados e operar funções de descoberta como identificação de aglomerados, regressão e identificação de padrões;
- **Visualização dos dados:** é o componente responsável por gerar representações visuais dos dados da forma como já foi definido neste texto;
- **Usuário:** é o responsável por guiar o processo de descoberta através da formulação de hipóteses, da verificação das mesmas e do delineamento de conclusões.

No modelo proposto, o usuário deve ser beneficiado pelos recursos promovidos por cada um dos componentes anteriormente mencionados para alcançar seus objetivos. Atualmente, a interação entre estes quatro elementos tem o usuário como peça central da operação; ele é o responsável por gerenciar a informação e aplicar ferramentas de análise e/ou visualização através de três *interfaces* distintas, como se observa na figura 3(a). Isto é, o analista não conta com uma *interface* única onde possa somar os recursos oferecidos por cada um dos componentes do sistema, ao invés disso é obrigado a operar um a um os diferentes softwares com suas peculiaridades.

O modelo descrito em (Keim, Lee et al. 1995) propõe a utilização do componente de visualização como *interface* para os sistemas de gerenciamento e análise, conforme é apresentado na figura 3(b). É argumentado que o eficiente sistema de comunicação, promovido pela interação visual entre computador e ser humano, habilita o usuário a gerenciar todos os componentes através de uma única *interface*. Desta forma, o analista de dados não é envolvido por várias ferramentas, mas ao contrário, através da utilização de apenas uma delas pode se especializar e tirar o melhor proveito de todo sistema.

Neste modelo de interação é suposto que todos os três domínios (análise estatística, consulta a banco de dados e visualização) estejam completamente integrados, o que, segundo (Keim, Lee et al. 1995), é necessário para uma exploração eficiente e efetiva.

O projeto e desenvolvimento de um sistema com a magnitude do modelo acima descrito não estão no escopo do presente trabalho. No entanto, a contribuição da referida proposta, aqui avaliada, deve ser considerada para elaboração de sistemas de visualização objetivando resultados

que possam ser integrados entre si, de maneira que permitam compor uma ferramenta robusta com potencialidades analíticas, visuais e de gerenciamento segundo o modelo estudado.

O software, que será apresentado no decorrer do capítulo 4 do presente trabalho, é um trabalho inicial desenvolvido no GbDI seguindo esta linha, entre suas funcionalidades incluem-se a capacidade de consulta de dados e análise estatística acopladas à cena de visualização dos dados em análise.

Figura 3 - Modelo de interação extraído de (Keim, Lee et al. 1995). (a) A arquitetura atual tendo o usuário como componente central. (b) A arquitetura proposta tendo a visualização como componente central.

3.5 - Técnicas de Visualização

Uma imensa variedade de esquemas de apresentação dos dados graficamente já foram propostas até agora. São sucessivas tentativas de se encontrar o "ovo de Colombo" da visualização, uma técnica que englobe todo o potencial investigativo desejado pela ciência de MD como já levantado nesta dissertação. Mas, independente do sucesso de um método ou de outro, esta pletora de idéias só pode ser mais bem compreendida através de um tratamento sistemático, isto é, uma classificação que permita agrupar técnicas semelhantes e organizá-las por suas vantagens e desvantagens, comparativamente. Nesta seção serão vistas duas classificações das técnicas de visualização, e na seção seguinte serão apresentados, com detalhes, exemplos elucidativos das classes revistas.

A classificação de Keim (Keim 2002) divide as técnicas em: projeções 2D/3D convencionais, baseadas em projeções geométricas, baseadas em ícones, orientadas a *pixels*, e em técnicas hierárquicas, além da possibilidade de combinação destas técnicas, o que se denomina técnica híbrida. Tais técnicas são resumidas a seguir:

- **Projeções 2D/3D convencionais:** abrangem um grande número de técnicas mais simples e amplamente utilizadas como plotagem em planos e espaços, gráficos de barras, *pie-charts*, *line graphs*, etc;
- **Técnicas baseadas em projeções geométricas:** têm como princípio o mapeamento de dados multidimensionais para padrões bidimensionais através da utilização dos valores presentes na base de dados como parâmetros para a geração de formas geométricas. Estas formas devem ser tais que o conteúdo da informação representada possa ser percebido e analisado visualmente em suas propriedades gráficas, sendo que quanto mais propriedades puderem ser percebidas individualmente, mais atributos dos dados serão discriminados. Como exemplos temos as Coordenadas Paralelas (Inselberg and Dimsdale 1990), as *Star Coordinates* (Kandogan 2000) e os *Scatter Plots* como descrito em (Ward 1994);
- **Técnicas baseadas em ícones:** onde cada item de informação é representado como um ícone, cuja aparência deve ser familiar ao ser humano, para que os atributos das entidades gráficas possam ser prontamente associados aos itens de dados em análise. Segundo (Pickett and Grinstein 1988), cor, forma e textura são características amplamente exploradas no design dos ícones, pois podem ser utilizadas simultaneamente sem perda de informação. Como exemplos temos as clássicas Faces de Chernoff (Chernoff 1973), os *Star Glyphs* (Chambers, Cleveland et al. 1983) e as *Stick Figures* (Pickett and Grinstein 1988);
- **Técnicas orientadas a *pixels*:** a idéia aqui é exibir cada atributo de um dado multidimensional através de *pixels* do dispositivo de exibição, um *pixel* para cada atributo fazendo uso de cores para representar os valores dos dados. É também calculado um fator, denominado Fator de Relevância (Keim and Kriegel 1994), baseado no qual os elementos serão ordenados para apresentação. Cada dimensão é apresentada em uma janela individual onde os elementos são comparados em relação a um atributo específico. A visualização pode ser gerada sobre todos os elementos de dados ou sobre

um subconjunto especificado por um consulta (figura 4). Em (Keim 2000) são apresentados os fatores a se considerar na construção de visualizações deste tipo: o arranjo dos *pixels* nas janelas, o mapeamento da cor, e o formato das janelas. Em (Keim and Kriegel 1996) e (Keim 2000) tem-se uma minuciosa abordagem do tema, em (Keim and Kriegel 1994) são apresentadas inúmeras variações da idéia básica, e em (Keim, Hao et al. 2001) é apresentada a idéia dos *Pixel Bar Charts* seguindo esta mesma linha de apresentação;

- **Técnicas hierárquicas:** nesta abordagem, o espaço k-dimensional é subdividido e os subespaços resultantes são apresentados de forma hierárquica, como por exemplo, na técnica denominada *Dimensional Stacking* (LeBlanc, Ward et al. 1990) que apresenta bidimensionalmente as dimensões em sucessivos níveis hierárquicos. Vale notar que hierarquia, aqui neste contexto, não possui conotação de importância ou subordinação, apenas ordem e detalhamento. (Shneiderman 1992) apresenta uma técnica que organiza os dados hierarquicamente e os apresenta através de uma imagem composta de retângulos cujos tamanhos e posições indicam a localização dos dados dentro da hierarquia.

Figura 4 - (a) Técnica de visualização orientada a *pixels* baseada em uma consulta sobre uma base de dados de cinco dimensões. (b) Uma alternativa de arranjo para apresentação de todos os atributos em uma única janela, como visto em (a).

O trabalho descrito em (Grinstein and Ward 2002) discute a classificação das técnicas de visualização por um ponto de vista bastante diferente, propondo que as técnicas podem ser classificadas dentro das seguintes caracterizações: geométricas ou simbólicas, utilizarem estímulos 2D ou 3D, ou se a exibição é estática ou dinâmica. Uma dada técnica teria sua classificação realizada sobre estes três quesitos simultaneamente.

Nesta abordagem, as técnicas **geométricas** baseiam-se em dados com vários atributos que podem ser mapeados para eixos escalados que constituam a exibição, como nas Coordenadas Paralelas e no *Scatter Plots*. As visualizações **simbólicas** visam representar os dados através de *pixels*, ícones, *arrays*, ou grafos cuja topologia tem grande importância na compreensão. Ambas as categorias, geométrica e simbólica, podem utilizar estímulos **2D ou 3D** de acordo com a noção espacial expressada pela cena, ou ambos simultaneamente (estereoscópico). Por último, as técnicas

podem ser rotuladas como **dinâmicas ou estáticas** segundo as possibilidades de interação apresentadas pela cena.

Observa-se a partir dos trabalhos apresentados em (Keim 2002) e (Grinstein and Ward 2002) uma grande divergência entre as propostas. A primeira tenta categorizar todas as técnicas de visualização dentro de apenas quatro classes, uma estruturação bastante simples, mas muito ambiciosa, cuja aplicação pode ser questionada, ou o conceito estendido. A segunda propõe oito possibilidades de um sistema de classificação cujas descrições das classes carregam uma quantidade maior de informações a respeito das características das técnicas. Por exemplo, é mais esclarecedor dizer que uma técnica é simbólica 2D e dinâmica do que categorizá-la apenas de orientada a ícones. Outra observação relevante é que ambas as propostas apresentam uma certa inadequação em suas nomenclaturas. Por exemplo, o termo "geométrico" torna-se impreciso quando, inevitavelmente, pensa-se em um *pixel* como um ponto, que é a mais primitiva das entidades geométricas, mas que, segundo as classificações estudadas, não determina técnicas com natureza geométrica.

Nesta breve discussão fica evidente que a questão sobre a classificação das técnicas de visualização ainda está aberta, haja visto que recentes proposições da literatura ainda não satisfazem o problema proposto. Ainda assim, as contribuições dos autores revisados são de grande valor ao servirem como subsídios para a elaboração de uma taxonomia mais completa, e ao poderem ser amplamente utilizadas neste estágio prematuro da Infovis.

3.6 - Exemplos de Técnicas de Visualização

Nesta seção busca-se apresentar vários exemplos de técnicas de visualização para que o leitor possa perceber mais claramente, do que se trata toda a teoria vista até aqui. Nesta etapa do texto, portanto, estamos utilizando o princípio da Visualização de Informações, ou seja, apresentar imagens para elucidar informações que, mesmo se revistas à exaustão em seu formato descritivo, não conseguem transmitir muitas das idéias que lhe são intrínsecas. Portanto, após toda a informação colocada até aqui, as imagens que serão agora apresentadas têm como intuito fixar alguns dos conceitos desenvolvidos, bem como de estimular o leitor a refletir e absorver novas descobertas que vier a fazer.

A seguir, será apresentada uma coleção de técnicas julgadas relevantes seja por sua representatividade na área de Visualização de Informações, por servirem de exemplos das classes vistas na seção anterior, pela qualidade reveladora de suas cenas, ou por serem parte integrante do presente trabalho, como será detalhado na seção 4.3. As técnicas serão exemplificadas sobre duas bases de dados amplamente conhecidas na literatura da área, a saber:

- Íris: possui 150 registros de cinco dimensões descrevendo três gêneros de uma flor cuja espécie é denominada Íris;
- Carros: possui 406 registros de oito dimensões com dados das características de carros fabricados entre 1970 e 1982 nos EUA, Europa e Japão. Ambas as bases podem ser conseguidas em <http://www.ics.uci.edu/AI/ML/MLDBRepository.html>.

Seguem-se as seguintes técnicas de visualização: Coordenadas Paralelas, *Scatter Plots*, *Star Coordinates*, *Table Lens*, *Star Glyphs* e *Dimensional Stacking*.

- Coordenadas Paralelas: proposta em (Inselberg 1985) e (Inselberg and Dimsdale 1990) é uma das técnicas mais referenciadas na literatura, como se observa em (Miller and Wegman 1991), (Wegman and Luo 1997), (Fua, Ward et al. 1999), (Siirtola 2000) entre outras referências. A técnica mapeia o espaço k -dimensional para as duas dimensões da tela utilizando k eixos equidistantes paralelos a um dos eixos do dispositivo de exibição. Os eixos correspondem às dimensões e são linearmente escalados de acordo com o menor e o maior valor da dimensão em questão, conforme é apresentado na figura 5, que foi gerada com o auxílio da ferramenta GBDIView¹. Cada item de dado é apresentado como uma linha poligonal interceptando cada um dos eixos no valor correspondente (Keim and Kriegel 1996). A grande desvantagem da técnica está no fato de que as linhas poligonais se sobrepõem, reduzindo drasticamente a quantidade de dados que podem ser visualizados.

¹ A ferramenta GBDIView reúne, em um único programa, os quatro componentes de software que suportam as técnicas de visualização utilizadas no projeto, como será descrito na seção 4.3.

Figura 5 - Coordenadas Paralelas, na cena uma filtragem interativa do conjunto de dados **Carros** de origem japonesa (verde) e européia (azul) que possuem quatro cilindros. Gerado com auxílio da ferramenta GBDIView.

- *Scatter Plots*: ou matriz de *Scatter Plots*, é uma das mais antigas e mais usadas técnicas para se projetar dados de alta dimensionalidade em duas dimensões. Nesta técnica as dimensões são combinadas duas a duas e plotadas uma em função da outra em uma grade de imagens, conforme é ilustrado na figura 6. A técnica é especialmente capaz de expressar correlações entre as dimensões da base de dados. Muitas variações da proposta podem ser encontradas na literatura, como os *HyperSlices* (van Wijk and van Liere 1993) e o *HyperBox* (Alpern and Carter 1991), que se baseiam, ambos, em explorar as propriedades funcionais das dimensões da base de dados.
- *Star Coordinates*: apresentado em (Kandogan 2001), e também em (Hoffman, Grinstein et al. 1997) onde é denominado *Radviz*. Nesta técnica n eixos correspondentes às n dimensões dos dados são projetados tendo como origem um ponto comum. Em seguida, cada item de informação é projetado como um ponto cuja posição é definida pela soma dos produtos dos vetores unitários dos eixos pelos correspondentes valores dos atributos, como se observa na figura 7(c).

Figura 6 - Scatter Plots com Link & Brush: a base de dados de carros exibida com a mesma seleção realizada na visualização das Coordenadas Paralelas na figura 5. Em destaque a relação "milhas por galão x peso" dos carros japoneses (verde) e europeus (azul) que possuem quatro cilindros. Gerado com auxílio da ferramenta GBDIView.

Esta técnica apresenta vantagens quando utilizada para apresentação de dados estatísticos (figura 7(a) e 7(b)), e para a definição de *scatter plots* de duas e três dimensões através da manipulação dos eixos. Já sua apresentação padrão com a projeção de todos os pontos sobre todos os eixos é pouco esclarecedora e ambígua, pois vários registros de dados podem ser projetados sobre o mesmo ponto, isto é, aglomerados apresentados visualmente são falsos. Um aperfeiçoamento mais esclarecedor da técnica exhibe, em conjunto com a distribuição dos pontos, a distribuição populacional das dimensões, esta técnica é denominada *Polyviz* e sua descrição pode ser encontrada em (Grinstein, Trutschl et al. 2001).

Figura 7 - Visualização de dados estatísticos do conjunto Iris através do *StarCoordinates* sem os pontos (a), e com os pontos (b). O eixo da dimensão *class* não está ativado em nenhuma das cenas, aqui ele é utilizado apenas para realização de filtragem interativa, sendo que vermelho representa a classe versicolor, verde a classe virginica e azul a classe setosa. Em (c) é descrito o cálculo da posição dos pontos no plano cartesiano. Ilustração gerada com auxílio da ferramenta GBDIView.

- *Table Lens* (Rao and Card 1994): esta é uma técnica extremamente intuitiva que ilustra a grandeza dos valores dos atributos dos dados de forma comparativa em relação a todo o conjunto de dados. Através da distorção interativa, apresentada pela técnica, o usuário é capaz de observar todo o volume de dados ao passo que concentra sua atenção em elementos de maior interesse. É uma técnica de fácil aprendizagem que em pouco tempo pode revelar correlações (figura 8) e elementos discrepantes, principalmente.

Figura 9 - *Star Glyphs* exibindo a base de flores da espécie Iris. Em vermelho temos o gênero setosa. Visualização gerada com auxílio da ferramenta XMDV (Ward 1994).

- *Dimensional Stacking*: utiliza eixos perpendiculares nos quais as dimensões são mapeadas uma dentro da outra. Na figura 10 é possível observar o esquema desta técnica. Nos eixos x e y são mapeadas as dimensões em dois níveis, no primeiro estão o comprimento e a largura da sépala, e no segundo, mais interno, estão o comprimento e a largura da pétala dos registros do conjunto de dados Íris.

Figura 10 - Esquema hierárquico da técnica *Dimensional Stacking*, gerada com auxílio da ferramenta XMDV (Ward 1994), o destaque em vermelho representa o gênero setosa da espécie Iris que possui sépalas mais largas e estreitas, e pétalas menores que as demais. Em verde temos os gêneros versicolor e virginica.

O número de técnicas de visualização tem crescido incessantemente nos últimos anos, sendo que uma análise de todas elas é inviável. Contudo, acredita-se que a apresentação acima, com as técnicas selecionadas, seja capaz de ilustrar a ciência da Visualização de Informações, seu potencial, aplicabilidade e abrangência. Em (Hoffman and Grinstein 2002) uma revisão mais abrangente das técnicas de visualização pode ser encontrada. Deve-se esclarecer também que não há uma visualização que possa ser considerada a melhor de todas, ao invés disso, percebe-se que há uma certa vantagem de uma ou outra técnica para dados com características específicas. Em determinadas condições as virtudes de uma técnica são amenizadas e em outras suas qualidades são mais notáveis. Algumas abordagens são qualificadas para melhor apresentarem *aglomerados* em outros casos para melhor apresentarem correlações, e assim por diante; mas em geral todas transmitem uma percepção global da base de dados que a suporta. A quantidade de registros exibidos também influencia diretamente nas possibilidades de um tratamento ou de outro. Isto se

deve à sobreposição de elementos de visualização e/ou da área necessária para a exibição destes elementos, que podem fazer com que os limites dos atuais dispositivos de exibição sejam rapidamente alcançados.

Portanto, não há uma técnica que possa satisfazer todas as necessidades, que contemple um poder de investigação tal que as demais abordagens possam ser desprezadas. De fato, verifica-se que as técnicas complementam-se e se interligam, por exemplo, através da técnica de *Link & Brush*, enquanto que isoladas são pobres e pouco auxiliam na exploração dos dados. Todos estes fatores apontam para a necessidade de disponibilização de múltiplas visualizações capazes de, juntas, constituírem um ambiente de exploração visual mais completo, onde o usuário possa utilizar vários meios para gerar conhecimento a partir de dados.

3.7 - Desempenho no projeto de técnicas de visualização

Segundo (Siirtola 2000), a interação com as técnicas visuais de MD deve ocorrer segundo o princípio da "manipulação direta", que pode ser entendido como a habilidade de interagir com uma cena de visualização de tal forma que, a reação a uma ação do usuário ocorra dentro de um limite de tempo suficiente para que seja estabelecida a correlação entre o que ocorreu na cena e a ação do operador. Conhecido como limite de causa e efeito (Card, Mackinlay et al. 1999), este tempo é de aproximadamente 0.1 segundo, que é o tempo máximo antes que ação e reação se tornem eventos disjuntos. O princípio da "manipulação direta" certamente deve ser satisfeito na implementação de técnicas de Visualização de Informações, de outra forma, o usuário a frente de um sistema notadamente lento, com reações retardadas, não será capaz de extrair qualquer proveito da ferramenta. De fato, o usuário tende a abandonar a utilização de tal sistema em um curto período de tempo, o suficiente para esgotar sua complacência (Shneiderman 1984).

O desempenho computacional dos algoritmos que constituem a ferramenta de visualização é o fator cabal do qual depende a satisfação do princípio da "manipulação direta". E, naturalmente, o desempenho de um sistema desta natureza é função da quantidade de dados sendo analisada e do poder computacional disponível, como exibido na equação 1. Por poder computacional entende-se o potencial do conjunto que define a capacidade de um sistema composto por dispositivos de armazenagem em massa, memória, processador e largura de banda

para transmissão/recebimento de dados. Independentemente do poder computacional presente, o objetivo de desenvolver algoritmos otimizados, que não sobrecarreguem a demanda por memória e processamento, deve ser constantemente perseguido por projetistas e desenvolvedores.

$\text{coeficiente de desempenho} = \text{poder computacional} / \text{quantidade de dados} \quad (1)$
--

A cada novo passo de interação com um sistema visual, normalmente, é necessário refazer os cálculos para regerar a projeção no dispositivo de exibição. Isto se deve à necessidade de reposicionamento dos objetos, de re-definição de suas cores e mesmo de recomputação de sua apresentação para gerar um novo elemento visual. Esta carga de processamento, inerente à visualização computacional, dá-se graças à dinâmica prevista pela interação visual, que determina que os mecanismos de exibição de dados têm a tarefa de refazer a imagem sendo exibida ao usuário, geralmente de alta resolução. Portanto, deve ser preocupação do projeto de software de visualização que a recomputação seja minimizada reservando os recursos para refazer a exibição apenas quando, de fato, esta estiver desatualizada por ação da atividade do usuário.

A memória, uma preocupação de qualquer sistema computacional em virtude dos ainda ineficientes sistemas de armazenamento em massa, é outro fator que merece consideração. Geralmente, a exploração visual de dados combina diversas técnicas que, espera-se, possam ser somadas para que o usuário alcance melhores resultados a partir de um sistema heterogêneo de investigação, isto é, onde indícios de conhecimento possam ser colecionados de diversas fontes. É preciso, para que o sistema não se sobrecarregue, que a combinação de diversas técnicas visuais não se traduza em um consumo proporcional de memória. Para tanto, os dados que são fonte a partir da qual as cenas serão produzidas, devem ser alocados em um espaço de memória comum, sendo acessados de maneira otimizada através de uma *interface* de programação, a qual deve ser disponibilizada às unidades responsáveis pelo traçado das diferentes cenas. Este espaço de memória comum também deve ser utilizado para possibilitar a integração prevista pelo conceito de *Link & Brush*, propagando as redefinições de exibição prescritas pelo usuário às demais cenas.

Mesmo respeitando-se rígidas diretivas de projeto de software, o desenvolvimento de ferramentas visuais se depara com um dilema. Pelo fato, já apontado, de que o desempenho de um sistema de visualização é função do poder computacional disponível e da quantidade de dados em

análise, pode-se concluir que jamais um sistema será pleno em satisfazer quaisquer necessidades de investigação ao mesmo tempo em que proporciona um ambiente respeitando o conceito de "manipulação direta". Isto se deve ao fato de que a quantidade de informações coletadas cresce em ordem geométrica ao passo que o poder computacional disponibilizado comercialmente não aumenta no mesmo passo. Segundo (Fayyad 1998), a capacidade de armazenar cada vez mais dados decorre de dois grandes fatores: o barateamento dos custos dos equipamentos computacionais de armazenagem (discos e memória), e do crescimento da utilização de SGBD's como centros da organização de informações. Estes últimos beneficiados por processadores mais poderosos e por redes de computadores mais rápidas. Em (Munzner 2002) afirma-se que a velocidade dos futuros processadores irá, sem dúvidas, continuar a avançar de acordo com a Lei de Moore², mas a quantidade de dados para serem processados irá crescer ainda mais rapidamente.

Em (Keim and Kriegel 1995) é afirmado que com a disponibilização de tecnologias para coletar e armazenar dados, a lacuna entre a quantidade de informações que deveria ser visualizada e a quantidade que pode ser visualizada está crescendo. Ou seja, a equação 1 tende a zero quando o poder computacional tende ao infinito.

No entanto, apesar deste cenário pessimista, na prática pode-se afirmar que existe uma quantidade máxima de dados que podem ser analisados, ou mesmo coletados, ao mesmo tempo em que conclusões úteis possam ser geradas a partir deles, por mais poderoso que seja o sistema de análise. Na verdade, a informação sempre poderá ser particionada, filtrada, pré-processada e transformada de acordo com as etapas previstas pela teoria de KDD, diminuindo ou distribuindo a carga sobre o processamento disponível. Além disso, apesar das abstrações da formalização matemática da equação 1, a capacidade e o interesse em se adquirir massas de dados são limitadas. Portanto, o sistema ideal que contempla a absoluta eficiência na tarefa de interação pode, sim, ser alcançado respeitando-se os preceitos básicos de projetos de software.

²A lei de Moore, enunciada em 1965 por Gordon Moore, então presidente da fábrica de microprocessadores Intel, estabelece que o poder computacional dos microprocessadores dobra a cada 18 meses.

3.8 - Limitações das técnicas de visualização

Uma limitação bastante conhecida da Visualização de Informações está na quantidade de elementos que podem ser apresentados simultaneamente. Em (Rundensteiner, Ward et al. 2002), os autores afirmam que técnicas convencionais de visualização multivariada não apresentam boa escalabilidade com respeito ao número de itens do conjunto de dados, resultando em uma cena com inaceitável grau de desordem. Segundo (Keim and Kriegel 1996), o número aproximado de elementos que a técnica das Coordenadas Paralelas é capaz de apresentar gira em torno de 1000 elementos. Com efeito, as técnicas geométricas em geral, baseadas em projeções bidimensionais ou tridimensionais, rapidamente alcançam o limite do que pode ser considerado compreensível. Isto se deve à sobreposição dos elementos de dados que foram mapeados para uma mesma, ou uma muito próxima, coordenada da cena de visualização. Com o acúmulo de itens em determinadas áreas da imagem, a visualização passa a apresentar "borrões", regiões totalmente preenchidas de cor ao mesmo tempo em que estão rodeadas de elementos mais suaves como linhas e pontos.

Mas este problema não é exclusivo das técnicas baseadas em projeções, isso ocorre de maneira ainda mais grave nas técnicas orientadas a ícones, que esbarram em um limite bem menor devido ao tamanho superior de seus elementos gráficos. Já as técnicas orientadas a *pixels* (Keim 2000) alcançam uma escalabilidade bem maior já que se baseiam no menor tamanho possível para um elemento gráfico: um único *pixel* por dado sendo apresentado. No entanto, estão restritas aos limites do dispositivo de projeção, o que gira em torno de 2 milhões de *pixels* ou, para um conjunto com n dimensões, a n -ésima fração deste número. Por sua vez, as projeções hierárquicas também não trazem muito alívio ao problema de escalabilidade. A visualização destas técnicas rapidamente deixa de ser compreensível com o aumento do número de dimensões. Em (Keim and Kriegel 1995) é afirmado que mesmo se fosse possível exibir um dado por *pixel* do dispositivo, ainda assim a quantidade de informações que poderiam ser visualizadas é bastante limitada.

Em (Fua, Ward et al. 1999) é apresentada uma solução, com elevado custo de processamento, que se baseia no agrupamento hierárquico dos dados de forma dinâmica. Em (Wong and Bergeron 1995) é feita uma abordagem baseada em níveis de transformadas de *wavelets*, que proporciona a observação da cena com diferentes graus de detalhamento. Em (Miller and Wegman 1991), com a técnica denominada *density plots*, a saturação das cores das

Coordenadas Paralelas é determinada em função das densidades de probabilidade dos dados sendo visualizados, sendo que as maiores densidades determinam linhas mais visíveis. Em (Wegman and Luo 1997) os *density plots* são discutidos e uma implementação é apresentada.

O problema de escalabilidade das técnicas de visualização é discutido no capítulo 4 deste documento onde uma abordagem baseada no grau de sobreposição dos dados é formalizada.

3.9 - Considerações Finais

Neste capítulo foram apresentados os conceitos e os objetivos da Visualização de Informações, uma das subáreas da Visualização de Informações Auxiliada por Computador, enquanto que o outro ramo desta disciplina, a Visualização Científica, também foi abordado para a adequada discriminação dos dois campos da Infovis. Prosseguindo com o tema de análise visual de dados, foram apresentadas algumas técnicas de pré-processamento de dados capazes de melhorar as visualizações resultantes.

Em seguida foi visto o tópico de interação, um dos mais importantes relacionados à Visualização de Informações, em conjunto com duas propostas de classificação de técnicas desta natureza, além de um modelo para desenvolvimento de sistemas de visualização/banco de dados com alto grau de interatividade.

Foram ainda apresentadas duas propostas de classificação de técnicas de visualização de dados seguidas de exemplos ilustrativos dos conceitos e dos resultados alcançados nesta área de pesquisa. Por último foi feita uma breve discussão sobre questões relacionadas ao desempenho e à limitação de sistemas de visualização que contemplem interatividade.

Capítulo 4 - O Projeto Desenvolvido

4.1 - Considerações Gerais

O presente trabalho pode ser dividido em duas principais frentes de atividades, a saber:

- o desenvolvimento de um Arcabouço para a elaboração da visualização das características extraídas das imagens organizadas pelo sistema cbPACS (Bueno 2002) em desenvolvimento pelo GBDI-USP, permitindo a percepção de agrupamentos e de seu inter-relacionamento, além de apontar os chamados elementos de exceção;
- propostas de algoritmos de apresentação dos dados de forma visual para domínios multidimensionais e métricos, permitindo uma boa compreensão da informação obtida.

O Arcabouço foi desenvolvido através da elaboração de componentes de software que agregam os algoritmos desenvolvidos para a geração das visualizações. Neste contexto, diversos melhoramentos, capazes de alavancar os resultados da utilização de técnicas visuais, foram propostos e implementados a fim de permitir a exploração visual das características extraídas das imagens do sistema cbPACS . O trabalho final concluiu uma versátil e poderosa ferramenta de visualização que permite a detecção de aglomerados e tendências através da integração de instrumentos visuais e estatísticos em um *Pipeline* de visualização, como será descrito nas próximas seções.

4.2 - A Ferramenta FastMapDB

O Grupo de Bases de Dados e Imagens do Instituto de Ciências Matemáticas e Computação da USP de São Carlos - GBDI-USP - é formado por pesquisadores cujas áreas de interesse concentram-se na utilização de bases de dados para a organização e recuperação de

imagens e dados complexos. Os temas de pesquisa incluem processamento de imagens, estruturas de indexação, bancos de dados, recuperação de imagens por conteúdo, entre outros. Nota-se, portanto, que as atividades do grupo estão inseridas em dois temas: bancos de dados, como aplicação gerenciadora de informações de diversas naturezas, e imagens, cujo tratamento e utilização apropriados podem trazer grandes benefícios.

Assim, dentro do GBDI-USP um projeto inicial de visualização de dados foi proposto em (Traina, Caetano et al. 2001), que apresentou a ferramenta FastMapDB, constituída de técnicas de pré-processamento de dados (como descrito na seção 3.3), técnicas de interação (seção 3.4) e técnicas de projeção simples (seção 3.5), para a análise visual de conjuntos de dados. É a partir desta ferramenta que o projeto aqui apresentado baseia seu principal eixo de trabalho, razão pela qual esta seção é dedicada à descrição da ferramenta FastMapDB.

Estendendo o trabalho apresentado em (Traina, Caetano et al. 2001), o presente projeto visa a inserção de técnicas mais sofisticadas de Visualização de Informações como ferramentas para exploração de bancos de dados. Assim, uma das contribuições do projeto aqui desenvolvido vem do casamento entre a ferramenta FastMapDB e o ferramental de visualização já proposto na literatura. A abordagem minuciosa desta união será feita na seção 4.11, após a revisão da contribuição de todo trabalho, enquanto que os detalhes da ferramenta FastMapDB serão fornecidos aqui como suporte para a próxima seção.

A ferramenta FastMapDB é baseada no algoritmo proposto em (Faloutsos and Lin 1995), que possibilita a redução de dimensionalidade de um conjunto n -dimensional para um conjunto k -dimensional onde $k < n$. O mecanismo do algoritmo funciona buscando distribuir, entre as dimensões originais da base de dados, a inevitável distorção que decorre da redução dimensional. O algoritmo baseia-se na Lei dos Cossenos da geometria e tem complexidade linear, o que lhe confere um ótimo desempenho e escalabilidade.

Figura 11 - O FastMapDB e seus vários elementos de interação e análise.

No FastMapDB o algoritmo FastMap é utilizado de maneira que um dado conjunto de informações n -dimensional tenha seu número de dimensões reduzido para 3, o que possibilita sua projeção em um espaço cartesiano. Como se observa na figura 11, o espaço, onde os pontos que representam a base de dados são projetados, é apresentado na forma de um cubo interativo (imagem à esquerda) no qual diversas propriedades dos dados podem ser identificadas. Na cena gerada pela ferramenta, tem-se ainda a projeção do cubo interativo em um plano cartesiano (imagem à direita), neste plano é possível realizar operações de zoom e rotação cujos efeitos da interação se observam inclusive na imagem do espaço tridimensional delimitado pelo cubo interativo. É possível também, no FastMapDB, a seleção de subconjuntos de pontos através de seleções visuais realizadas dentro do cubo de interação, deste modo os dados selecionados podem ser observados em seu formato original provendo um mecanismo de consultas com inúmeras possibilidades exploratórias.

Graças às propriedades do algoritmo FastMap, a visualização do espaço cartesiano, populado pelos pontos que representam a base de dados, é capaz de apresentar de forma bastante

eficiente algumas das principais características dos dados: sua distribuição global dentro do domínio do qual pertencem, elementos discrepantes e, principalmente, *aglomerados*. Sendo que estes aspectos são facilmente percebidos interagindo-se com a cena gerada pela ferramenta, o que leva o analista a identificar aglomerados e pontos discrepantes dentro do conjunto de dados.

A ferramenta FastMapDB permite gerenciar dados previamente classificados. Para tanto, um atributo (classificador) dos dados deve ser fornecido para que cada elemento de informação seja discriminado segundo uma categoria na qual se enquadram. Desta maneira, uma vez que a natureza das informações seja explicitada através do atributo classificador, a ferramenta utiliza esta informação para distinguir as classes presentes no domínio de dados através da cor e forma dos pontos projetados no espaço cartesiano, como ilustrado na figura 11. A utilização deste esquema de apresentação dos dados permite uma análise comparativa entre objetos de informação cuja espécie é previamente conhecida. Por meio deste tratamento as categorias do conjunto de dados podem ser verificadas e comparadas, enaltecendo propriedades comuns ou antagonismos potencialmente valiosos para a análise.

A utilização do algoritmo FastMap (Faloutsos and Lin 1995), (Traina, Traina et al. 1999), (Traina, Caetano et al. 2001) mostrou-se muito eficiente para análise de bancos de dados de diversas naturezas, classificados ou não. E sua utilização como base do presente projeto foi vislumbrada como uma proposta cujo maior proveito será a potencialização das faculdades exploratórias já comprovadas para análise de informações através do algoritmo FastMap.

4.3 - Duas contribuições à identificação visual de aglomerados na ferramenta FastMapDB

Uma das maiores virtudes da visualização gerada pela ferramenta FastMapDB é a possibilidade de identificação de aglomerados, que pode ser realizada de maneira bastante simples através da percepção de aglomerados nas regiões mais populadas do espaço cartesiano delimitado pelo cubo interativo da cena. Essa identificação tem grande potencial de utilização tanto em conjuntos de dados sem classes definidas como em conjuntos onde subconjuntos são previamente categorizados através de um atributo classificador. Em conjuntos não classificados, os aglomerados indicam tendências da distribuição global levantando evidências de que o conjunto possui classes ainda não determinadas. Já nos conjuntos onde as categorias já foram reconhecidas,

a visualização dos aglomerados lança uma nova dimensão de introspecção sobre os dados rotulados: a qualidade de uma classe pode ser expressa pelo quão compacto é o aglomerado que ela determina, o que pode ser facilmente identificado pelo sistema de visão humano, para o qual os elementos das diferentes classes são imediatamente discernidos com base no sistema de exibição da ferramenta.

No entanto, a percepção das características de uma classe pode ser prejudicada em algumas distribuições de dados. Os elementos das diversas classes podem se emaranhar dificultando a visualização correta dos limites dos aglomerados por elas determinados. Desta forma, visando um auxílio específico para este caso, isto é, a análise visual de classes através da perspectiva da presença de aglomerados, projetou-se um sistema cujo produto, de natureza quantitativa, é capaz de caracterizar o aglomerado determinado pelos elementos de uma dada classe. A qualidade de um aglomerado vem a ser o quão fechado estão os seus elementos dentro dos limites por eles próprios delineados, o que pode ser medido pela variância presente na distribuição destes elementos dentro do espaço que os engloba. Este espaço nada mais é do que o *Minimum Bounding Box* (MBB) que contém todos os pontos daquela classe, e que pode ser observado na visualização gerada pelo FastMapDB. Formalmente, tem-se:

Proposição 1: as coordenadas, que definem o subespaço tridimensional de um conjunto de pontos no espaço cartesiano, podem ser conseguidas pela identificação do maior e do menor valor em cada uma das 3 dimensões que definem o domínio espacial dos pontos. Os valores encontrados são as primitivas dos oito pontos que perfazem os vértices do MBB dentro do qual o conjunto de pontos está inserido.

Uma vez que o MBB de uma classe é conhecido, prossegue-se calculando a distância de cada ponto até o centro do paralelepípedo. A variância destas distâncias irá indicar o quão denso o subespaço em questão é e, por conseguinte, a qualidade do aglomerado definido. Por qualidade entende-se que quanto mais compacto um aglomerado é, mais similar é o comportamento dos pontos da classe que o definiu, e quanto maior a variância expressa pelos pontos, mais heterogeneidade está presente.

A segunda métrica elaborada como auxílio à validação de classes sobre a caracterização de aglomerados procura expressar a pureza dos aglomerados de dados obtidos. De espécie *booleana*, esta métrica indica se uma determinada classe tem seu correspondente MBB invadido por elementos de uma outra classe, isto é, se os elementos estão embaralhados no espaço

cartesiano. A importância desta métrica vem da confusão muitas vezes gerada por elementos que se misturam na visualização, impedindo a percepção visual acurada dos MBB's das classes em análise.

A utilização destes dois instrumentos agregados à ferramenta FastMapDB foi de especial importância para o desenvolvimento de um método para avaliação de técnicas de extração de características utilizadas para caracterização de imagens, o qual é descrito a seguir.

A pesquisa por técnicas de extração de características ocupa um grande número de pesquisadores que buscam meios mais eficientes de recuperação de imagens baseando-se em vetores de características capazes de identificar as imagens através de propriedades extraídas (*features*) das mesmas (Rui, Huang et al. 1999). Várias novas técnicas de extração são propostas todos os anos, e a avaliação da eficácia destas técnicas se baseia, geralmente, na construção de gráficos de *Precision and Recall* (Baeza-Yates and Ribeiro-Neto 1999), cuja metodologia de execução requer um exaustivo trabalho de avaliação da recuperação de imagens proporcionada pelo extrator. Utilizando a versão estendida do FastMapDB, propõe-se um método alternativo, e bem mais simples, de avaliação de técnicas de extração de características baseada nas métricas desenvolvidas.

Tal metodologia baseia-se em dois pressupostos:

- 1) Os vetores de características são capazes de caracterizar as imagens a partir das quais eles foram gerados, e;
- 2) A visualização de dados tem o poder de expressar as principais propriedades de um conjunto de dados que, de outra forma, não poderiam ser analisados em seu formato original por um ser humano.

A conciliação destes dois preceitos leva à dedução de que a visualização dos vetores de características é capaz de revelar o potencial de caracterização de imagens que será descrito por tais vetores, quando utilizados para a recuperação de imagens. O discernimento deste potencial baseia-se na informação visual proporcionada pela ferramenta FastMapDB e nas métricas de qualidade de aglomerados visuais que foram implementadas. Assim sendo, o potencial dos extratores de características deve ser proporcional à qualidade da discriminação de vetores extraídos a partir de imagens sabidamente semelhantes (classes), e que na ferramenta de visualização, espera-se, determinem aglomerados de boa qualidade dada a similaridade visual de seus dados geradores.

A validação da proposta foi feita aplicando-se a técnica a quatro diferentes espécies de extratores de características baseados sobre transformadas de *wavelets* utilizando diferentes abordagens para sumarização das imagens: média, entropia, energia (Albuz, Kocalar et al. 2001) e textura (Gabor (Lee 1996)), sendo que os detalhes dos extratores são apresentados em (Castañón 2003). Foram gerados vetores de características de sete grupos (classes) de imagens sabidamente semelhantes com 30 elementos cada um, e de mais 290 imagens distintas das demais, totalizando 500 vetores de características. Na figura 12, os resultados das visualizações dos vetores de características gerados por cada um dos extratores sobre os sete grupos de imagens, além das métricas calculadas pela ferramenta, podem ser confrontados com seus respectivos gráficos de *Precision and Recal*. A visualização obtida pela ferramenta corrobora os resultados das curvas de *Precision and Recal*, indicando o potencial da ferramenta para analisar a usabilidade de técnicas de extração de características.

4.4 - Técnicas de Visualização de Informações Utilizadas

A seleção e o número das técnicas utilizadas para a implementação do principal objetivo do presente projeto, a construção de um Arcabouço de visualização, tiveram como critérios a qualidade das técnicas e o tempo disponível para tal projeto. A princípio, as técnicas mais mencionadas na literatura foram prontamente escolhidas: as Coordenadas Paralelas e os *Scatter Plots*. Ambas com eficiências e limitações conhecidas e comprovadas, apresentaram-se como opções naturais de implementação, pois são técnicas capazes de promover um satisfatório contato inicial com a implementação de técnicas visuais graças à complexidade relativamente baixa e à eficiência de análise das cenas geradas.

As outras técnicas selecionadas: as Coordenadas Estelares e as *Table Lens* foram escolhidas graças à flexibilidade de apresentação aliada à natureza das técnicas, distintas das anteriores na forma de organizar os elementos de dados, com abordagens únicas de exploração. Estas duas últimas técnicas, integradas às demais, constituíram um conjunto heterogêneo bastante adequado ao projeto; em cada apresentação o usuário se depara com uma perspectiva nova e diferente das informações podendo adquirir uma maior introspecção em função deste fator. Abordagens 3D foram deixadas para última instância devido ao fato de que o salto em complexidade de implementações 2D para 3D não ser linear, tendo como única contribuição alcançada, seguindo esta linha, a possibilidade de projeção das Coordenadas Estelares no espaço

cartesiano. No entanto, com interação bastante reduzida. As 4 técnicas escolhidas já foram apresentadas na seção 3.6, aqui será realizada uma segunda análise abordando outros aspectos que não sua simples descrição.

As Coordenadas Paralelas são especialmente apropriadas para a identificação de características unidimensionais como densidades marginais, características bidimensionais como correlações e estruturas não lineares, além de características multidimensionais como aglomerados, hiperplanos e nós (Wegman and Luo 1997). As Coordenadas Paralelas perdem eficiência quando o número de elementos é muito alto, o que inevitavelmente acarreta a sobreposição de elementos gráficos que perdem a individualidade.

Figura 12 - Em (a) temos exemplares dos sete grupos de imagens utilizadas no experimento. (b) mostra a curva de *Precision and Recal* dos quatro extratores ao responder consultas por similaridade (vizinhos mais próximos). Em (c) temos a visualização dos vetores de características extraídos das imagens exemplificadas em (a). (d) mostra as métricas calculadas pela ferramenta FastMapDB.

Os *Scatter Plots* são altamente especializados na identificação de correlações entre as dimensões, sua matriz de gráficos determina o cruzamento de todas as dimensões, e na cena resultante esboços de funções tornam-se evidentes ao olho humano, que é capaz de

automaticamente traçar curvas de interpolação. A técnica perde eficiência com um número de dimensões muito alto (Hoffman and Grinstein 2002) quando a quantidade de gráficos cresce dificultando a observação.

As Coordenadas Estelares possuem uma característica especial que é seu alto potencial de interação. Os usuários podem habilitar e desabilitar eixos, alterar suas direções, realizar operações de escala sobre eles, recuperar o valor dos atributos visualmente, etc. Estas operações permitem ao usuário manipular os pontos que representam os dados de maneira bastante eficiente, proporcionando uma forte percepção da distribuição das informações. O usuário, através da manipulação da cena, pode determinar *Scatter Plots* de duas ou três dimensões com vantagens próprias desta técnica e pode confrontar atributos ao posicionar seus eixos em direções opostas. Além disso, como proposto nesta dissertação, é possível observar de maneira muito eficiente dados estatísticos, que são muitas vezes elucidativos por si só. Porém, um grande número de dimensões é a maior limitação da técnica, quando a interação pode se tornar maçante e ineficiente.

A técnica *Table Lens* determina um ambiente adequado à identificação de correlações entre atributos de dados. A possibilidade de ordenar a apresentação por um atributo específico a torna a mais intuitiva das técnicas aqui utilizadas, as correlações presentes nos dados apresentam-se quase que instantaneamente mediante manipulação do usuário. A técnica também possui grande potencial de apresentação de elementos discrepantes ao exibir de maneira bastante eficiente todos os atributos de forma comparativa. A técnica é limitada na identificação de aglomerados e na análise de dados que não possuam correlação entre as dimensões.

Técnica	Classificação (Keim)	Classificação (Grinstein)	Maior Virtude	Maior Limitação
Coordenadas Paralelas	Geométrica	2D, Geométrica Dinâmica	Aglomerados	Não escalável
Scatter Plots	Geométrica	2D, Geométrica, Dinâmica	Correlações	Número de Dimensões
Coordenadas Estelares	Geométrica	2D/3D, Geométrica, Dinâmica	Apresentação de Dados Estatísticos	Número de Dimensões
Table Lens	Geométrica	2D, Geométrica, Dinâmica	Correlações	Aglomerados

Tabela 1 - As técnicas utilizadas no projeto e suas características.

Pela comparação que pode ser realizada a partir da tabela 1, percebe-se que as técnicas implementadas possuem uma boa diversificação no que diz respeito às suas principais características de análise. Este fato é especialmente importante dado o alto grau de integração que foi alvo de todo o projeto, conseguiu-se, desta maneira, um ambiente em que as diferentes abordagens se complementam constituindo uma ferramenta bastante útil.

4.5 - Integração das Técnicas

Uma das maiores preocupações durante o projeto foi a possibilidade de integração, segundo o conceito de *Link & Brush*, das técnicas que fossem utilizadas. Através do *Link & Brush*, como já revisto na seção 3.4.1, os resultados de interação com uma dada técnica podem ser imediatamente observados durante a visualização da cena segundo outras técnicas suportadas por um ambiente comum. O processo de interação utilizado no projeto é baseado na filtragem interativa, através da qual consultas podem ser realizadas visualmente. Portanto, a integração das técnicas teve que levar em consideração as peculiaridades de cada apresentação para que a sensação de inter-relação entre as várias cenas ficasse clara, isto é, os dados selecionados e os elementos gráficos de interação deveriam ser imediatamente percebidos em todas as diferentes abordagens.

As quatro representações gráficas dos dados, conforme mencionadas na Seção 4.3, podem ser grosseiramente caracterizadas como: baseada em segmentos de retas, baseada em uma matriz de funções, baseada na noção espacial de um plano cartesiano, e baseada em colunas de segmentos de retas. Percebe-se, nesta listagem, que não há qualquer semelhança na estrutura de definição das cenas em qualquer par confrontado. Esta heterogeneidade na projeção dos elementos gráficos levou à necessidade de criação de mecanismos de exibição dos componentes de interação que apresentassem equivalência cognitiva ao mesmo tempo em que preservassem as escolhas realizadas pelo usuário. Portanto, para cada uma das técnicas foi criado um componente, denominado selecionador, capaz de personificar de maneira inteligível o *Link & Brush* da filtragem seletiva objetivada.

Como se observa na figura 13, o componente selecionador das Coordenadas Paralelas é composto por segmentos de reta delimitados por setas na mesma direção vertical, mas com sentidos opostos sobre os eixos verticais; nos *Scatter Plots* os selecionadores só podem ser

definidos na diagonal principal onde os gráficos são funções de um mesmo atributo. As seleções são definidas por retângulos cujas largura e comprimento delimitam sempre o mesmo intervalo de dados, isto é, apesar de não possuírem mesmo tamanho, englobam uma porcentagem equivalente dos eixos vertical e horizontal, que são escalados pela mesma dimensão dos dados; as Coordenadas Estelares possuem selecionadores bastante simples na forma de segmentos de reta cuja direção é definida pelo vetor unitário do eixo sobre o qual estão definidos. Finalmente, as *Table Lens* possuem um selecionador bastante peculiar; ele é definido na forma de um retângulo posicionado sobre o cabeçalho das dimensões, e cujos lados verticais se projetam ao longo da coluna na qual fazem seleção. A grande particularidade deste meio de filtragem está nos objetos abrangidos pela consulta; apenas os elementos cujos extremos dos segmentos de reta estiverem dentro da lacuna definida pelo selecionador participam da apresentação.

Figura 13 - Os selecionadores utilizados para interagir com os componentes visuais. (a) Coordenadas Paralelas, (b) *Scatter Plots*, (c) Coordenadas Estelares, e (d) *Table Lens*.

O resultado da integração foi bastante satisfatório no sentido de propiciar uma análise mais completa dos dados. Na forma de uma instância comum de uma classe de programação, como será descrito adiante, as seleções realizadas em todas as n distintas visualizações ocupam a mesma área de memória. De fato, existe um único conjunto de seleções referenciado, enquanto que há n procedimentos de desenho destas seleções, fornecidos pelo desenvolvedor que se propuser a implementar uma nova técnica. A abordagem descrita proporcionou um modo de interação bastante rápido e integrado.

4.6 - Exibição Visual de Freqüências

Conforme discutido na seção 3.8, uma das limitações da Visualização de Informações está na quantidade de elementos que podem ser apresentados simultaneamente. Com este problema em mãos, buscou-se uma maneira para ponderar tal restrição através do desenvolvimento do que se denominou *Frequency Plot*, ou seja, a exibição visual de freqüências. Por freqüência se entende como o grau de sobreposição dos dados de um conjunto de informações em um determinado ponto. No trabalho proposto, esta idéia é implementada em quatro diferentes técnicas de visualização e a formalização matemática da técnica é apresentada em seguida:

Dado um conjunto arbitrário de valores $V = \{v_0, v_1, \dots\}$, uma função $q(v_i)$ que indica quantas vezes o elemento v_i aparece dentro do conjunto V ; e a função $m(V)$ que retorna a quantidade de repetições da moda do conjunto V . A freqüência de um valor $v_i \in V$ é dada por:

$$f(v_i, V) = q(v_i)/m(V) \quad (2)$$

Ou melhor, a função $f(v_i, V)$ fornece um número entre 0.0 e 1.0 que indica o quão freqüentemente um determinado valor v_i é encontrado dentro de um conjunto de valores V . Aqui neste trabalho procede-se aplicando a função f para cada uma das dimensões de um conjunto de dados armazenando-se o resultado para posterior utilização. Dado um conjunto C com n elementos e k dimensões, seus valores podem ser encarados como um conjunto com k subconjuntos de valores, um subconjunto para cada dimensão dos dados. Isto é, $D = \{\{D^0\}, \{D^1\}, \dots, \{D^{k-1}\}\}$, sendo que $|D^x| = n$. Partindo deste princípio, pode-se realizar o cálculo das freqüências, como já descrito, para qualquer elemento de um conjunto de dados. Isto é, dado um elemento k -dimensional $d_i = (d_i^0, d_i^1, \dots, d_i^{k-1})$ pertencente ao conjunto C , seu correspondente vetor k -dimensional de freqüências F_i será dado por:

$$F_i = (f(d_i^0, D^0), f(d_i^1, D^1), \dots, f(d_i^{k-1}, D^{k-1})) \quad (3)$$

Uma vez calculadas as freqüências dos dados do conjunto que se quer exibir, a idéia é expressar estes cálculos através de recursos visuais, nomeadamente cor e tamanho. Na implementação realizada, o valor da freqüência em um determinado ponto foi utilizado para

especificar a saturação da cor de seu representante visual. Sendo que cada uma das técnicas teve um tratamento especial em função de suas particularidades. Nas Coordenadas Paralelas um segmento de reta determinado por dois valores com diferentes frequências foi desenhado variando-se a cor segundo a interpolação dos valores das frequências de seus extremos. Nos *Scatter Plots*, além da variação de saturação, o tamanho do pontos representando os dados é proporcional à média das frequências dos dois valores que definem sua projeção. Na *Star Coordinates* é calculada a média das frequências para determinar a coloração e o tamanho dos pontos, no entanto os resultados não são muito esclarecedores. Na *Table Lens*, a frequência determina apenas a cor, sendo que nesta técnica valores iguais em uma mesma dimensão ordenada determinam um retângulo dentro da coluna de visualização, claramente os retângulos mais extensos possuem a coloração mais acentuada.

A visualização utilizando o recurso de frequência mostrou-se bastante eficiente para contornar o problema da sobreposição de elementos. Nas cenas resultantes a tendência global dos dados é imediatamente percebida na regiões mais notáveis da imagem, poupando ao analista um bom tempo que seria dispendido buscando-se as maiores concentrações dos elementos através de outros meios, como consultas em SQL. Ainda, na implementação aqui descrita há a possibilidade de visualização através das frequências de subconjuntos dos dados selecionados pela filtragem interativa. Esta possibilidade é muito útil na exploração de dados, pois proporciona uma análise minuciosa de partições dos dados que satisfaçam ao interesse de análise. Na figura 14 é exibida a visualização de uma base de dados contendo 457 registros com 11 atributos cada, representando os resultados laboratoriais de tecidos suspeitos de apresentar câncer. Nesta base existem 285 amostras de tecidos sadios (classe 0) e 172 amostras de tecidos doentes (classe 1). Na figura são contrastadas as visualizações convencionais e as visualizações que fazem uso do *Frequency Plot*, sendo que a caracterização das classes fica clara mesmo não se conhecendo o significado dos atributos apresentados na base de dados.

A limitação da exibição de dados por frequência está em bancos de dados onde os valores não se repetem, seja pela natureza dos dados ou por uma precisão muito alta dos valores dos atributos. Por exemplo, não é comum encontrar duas vezes um mesmo número real que tenha mais de cinco casas de precisão. Algumas alternativas para futuros trabalhos podem ser propostas como, por exemplo, considerar que dois valores são iguais aceitando-se uma determinada margem

de variação, sendo que esta margem pode ser definida computacionalmente através de uma análise prévia dos dados.

Figura 14 - Ilustração da técnica *Frequency Plot* sobre a base de dados de exames laboratoriais de tecidos para identificação de câncer. Em (a) é exibida a totalidade da base de dados segundo a frequência dos valores que a compõe. Em (b) e (c), respectivamente, podem ser observados os tecidos saudáveis (classe 0) e os tecidos doentes (classe 1) segundo a filtragem iterativa simples. Em (d) e (e) as mesmas seleções utilizando-se da técnica *Frequency Plot*. Uma breve análise das imagens (d) e (e) possibilita a caracterização dos tecidos saudáveis e dos tecidos doentes.

4.7 - Exibição de Dados por Relevância

Uma das técnicas de interação mais comuns, e provavelmente a mais importante, é a filtragem interativa que, como já descrito, permite a realização dinâmica de consultas visuais pelo usuário. Também denominadas de técnicas de *brushing* (realce), pois os dados selecionados ficam imediatamente destacados, geralmente, através de cores capazes de enfatizar a visualização de partes do conjunto de dados que sejam mais interessantes no contexto de uma aplicação. Dada sua importância, muitos trabalhos já trataram deste tema, como em (Ahlberg and Shneiderman 1994), (Keim and Kriegel 1995), (Martin and Ward 1995), (Siirtola 2000), e (Wong and Bergeron 1995).

As técnicas desenvolvidas para filtragem interativa são limitadas por se basearem exclusivamente em intervalos definidos sobre as dimensões visualizadas. Desta maneira, como em consultas baseadas em SQL realizadas em *interfaces* de texto de clientes de bancos de dados, os filtros definidos pelo usuário retornam apenas os dados que satisfazem plenamente às consultas. No entanto, a tarefa de análise poderia ser altamente potencializada caso tais consultas não fossem limitadas como descrito, ao invés disso, que fossem capazes de retornar, além dos dados categorizados, os dados que aproximadamente satisfizessem a consulta realizada, e o quanto a satisfizeram. Obviamente, esta possibilidade está além do potencial de clientes textuais, ao passo que ferramentas de visualização possuem um talento ainda inexplorado neste caráter.

Em (Keim and Kriegel 1996) é descrita uma técnica de visualização que se utiliza do potencial visual para exibição de resultados aproximados a consultas precisamente especificadas. No entanto, a abordagem é descrita como que se fosse exclusivamente adequada à visualização proposta naquele trabalho, que é baseada em janelas de *pixels* multicoloridos. Além disso, o processo desenvolvido naquele trabalho é carente de maior dinâmica interativa, as consultas devem ser realizadas e os resultados observados estaticamente. A geração de uma nova cena requer que todo o processo seja reiniciado a partir da *interface* de definição de consultas.

A proposta descrita nesta dissertação é denominada *Relevance Plot*, e descreve uma forma de agregar, a técnicas de visualização de qualquer natureza, a possibilidade de exibição de elementos de dados com diferentes graus de importância dada uma consulta de dados definida dinâmica e visualmente. Para tanto, assim como no *Frequency Plot* mostrado da seção anterior, atributos gráficos de forma e cor são os meios através dos quais uma dimensão a mais de

introspecção é fornecida ao analista de dados. A idéia é uma extensão da convencional filtragem interativa, com a diferença de que a interação utilizada não requer a definição de intervalos de valores nas dimensões da visualização, mas apenas um único valor em cada dimensão sobre a qual a definição dos dados de interesse se baseie.

O mecanismo do *Relevance Plot* requer que valores, ou pontos de relevância, pertencentes às dimensões sendo visualizadas sejam escolhidos. Além disso, em cada dimensão é definido um valor limite para que os dados sejam considerados relevantes, valores que ultrapassem este limite contribuem negativamente para o cálculo da relevância de um registro de informação. Desta forma, para um conjunto de dados n -dimensional, o qual assume-se que seus dados sejam processados na forma normalizada, isto é, todos os valores estão entre 0.0 e 1.0, temos:

Definição 1: o *Relevance Point* - RP (Ponto de Relevância) da i -ésima dimensão, ou RP^i , é o valor pertencente a esta dimensão que deve ser considerado para a determinação da relevância dos dados naquela dimensão. Apenas um RP pode ser escolhido por dimensão, sendo que as dimensões nas quais um ponto de relevância não foi definido têm RP igual a -1.

Note-se que não é necessária a escolha de um RP por dimensão, de fato, no mínimo um e no máximo n pontos podem ser definidos.

Definição 2: Seja R (*Relevants*) o número de RPs escolhidos pelo usuário. Então $1 \leq R \leq n$.

Uma vez que os pontos de relevância estejam definidos, os itens de dados devem ser analisados em relação ao que foi determinado. Para cada dimensão que teve um RP escolhido, prossegue-se com o cálculo da distância Euclidiana dos valores dos atributos dos dados até o ponto de relevância especificado naquela dimensão.

Definição 3: Para o j -ésimo elemento n -dimensional de informação $d_j = (x_j^0, x_j^1, \dots, x_j^{n-1})$, a distância de seu i -ésimo atributo até o i -ésimo RP é dada por:

$$D_j^i(x_j^i, RP^i) = |x_j^i - RP^i| \quad (4)$$

Deve-se lembrar que a distância máxima entre dois pontos quaisquer normalizados é 1.0.

Como já mencionado, para cada uma das dimensões uma distância máxima de aceitação é definida. Estes limites são chamados de *Max Relevance Distance - MRD* (Distância Máxima de Relevância), e são utilizados no cálculo da relevância total dos itens de dados.

Definição 4: A Distância Máxima de Relevância da i -ésima dimensão, ou MRD^i , é a distância máxima $D(x_j^i, RP^i)$ que o valor do i -ésimo atributo do j -ésimo item de dado pode assumir sem que contribua negativamente para o cálculo da relevância do elemento n -dimensional de informação. Os $MRDs$ assumem 0.1 como valores padrão, sendo que o usuário pode, opcionalmente, redefinir este limite. O MRD satisfaz $0.0 \leq MRD \leq 1.0$.

Baseando-se nos $MRDs$ e nas distâncias $D(x_j^i, RP^i)$ calculadas, valores denominados *Point Relevance Coefficient - PRC* (Coeficiente de Relevância de Ponto), são computados para os R atributos cujas dimensões serão consideradas no cálculo da relevância, ou seja, as dimensões que tiveram um RP definido.

Definição 5: O valor que determina a contribuição de um atributo de informação para o cálculo da relevância de um item n -dimensional de informação é denominado *Point Relevance Coefficient*. Para o atributo x_j^i , o PRC_j^i é dado por:

$$PRC_j^i = \begin{cases} 1 - \frac{(D_j^i(x_j^i, RP^i))}{MRD^i} & \text{se } D_j^i(x_j^i, RP^i) \leq MRD^i \\ -\frac{(D_j^i(x_j^i, RP^i))}{(1 - MRD^i)} & \text{se } D_j^i(x_j^i, RP^i) > MRD^i \\ 0 & \text{se } RP^i = -1 \end{cases} \quad (5)$$

A equação acima define que valores entre o ponto de relevância e sua respectiva distância máxima são graduados linearmente entre 1.0 e 0.0. Valores acima da distância máxima de relevância são graduados com coeficientes entre 0.0 para os mais próximos e -1.0 para os mais distantes. Os pontos das dimensões onde o RP não foi especificado são graduados com 0.0, de forma que não irão influenciar no cálculo final.

Com todos os PRC s calculados para todos os registros de dados, a análise de relevância de um item do banco de dados prossegue com o cálculo de um valor denominado *Data Relevance Coefficient* - DRC (Coeficiente de Relevância de Dado), que nada mais é do que a média aritmética dos PRC s computados para este dado específico. Para o j -ésimo dado de um conjunto de informações, o DRC_j é dado por:

$$DRC_j = \frac{\left(\sum_{i=0}^{k-1} PRC_j^i \right)}{R} \quad (6)$$

Onde R é o número de dimensões para as quais um RP foi definido. A figura 15 apresenta um exemplo de cálculo do DRC para um caso onde os RPs e os $MRDs$ foram definidos visualmente no ambiente das Coordenadas Paralelas.

Figura 15 - Exemplo de cálculo do DRC para um registro sendo exibido em uma cena das Coordenadas Paralelas.

Os testes de utilização do *Relevance Plot*, como se observa na figura 16, mostraram que a técnica é uma maneira eficiente de se realizar consultas aproximadas contribuindo muito para o trabalho analítico fundamentado sobre hipóteses. Devido à sua dinâmica e a seu caráter especulativo, a técnica possibilita ao analista definir novas hipóteses baseando-se na exibição de elementos gráficos que satisfazem aproximadamente seu interesse. Igualmente, a verificação das hipóteses também é altamente potencializada; através do conceito de relevância, o *feedback* das consultas não tem mais um caráter categórico, pelo contrário, a exibição dos dados proporciona uma percepção gradativa do comportamento dos dados, que podem satisfazer muito, pouco ou absolutamente nada a uma determinada consulta.

Figura 16 - Ilustração do *Relevance Plot*. Na cena temos a base de dados veículos sem o atributo de classe. Através da seleção especulativa proporcionada pela técnica é possível a identificação de duas tendências: em (a) veículos mais econômicos, menos potentes, mais leves e rápidos; em (b) veículos com menor desempenho e aceleração, mais potentes e pesados.

Observa-se ainda que as técnicas propostas acima são apenas o delineamento de um mecanismo mais poderoso. A grande motivação da idéia descrita está na possibilidade de utilização de diferentes métodos para o cálculo da distância entre os pontos que equivalem aos itens de dados. De fato, podemos utilizar funções matemáticas que capturem uma semântica mais complexa das informações para estabelecer a distância entre os pontos de dados e os pontos de relevância. A única exigência é que o resultado da análise seja normalizado, 0.0 para a relação mais fraca com variação linear até 1.0 para os pontos com relação mais forte. É possível ainda a definição de funções de distância diferentes em cada dimensão. Existe a possibilidade de definir mais de um ponto de relevância por dimensão e, por fim, é possível especificar pesos para os pontos de relevância e para as dimensões do conjunto de dados. Consultas extremamente complexas, que não podem ser realizadas utilizando-se o paradigma baseado em intervalos de dados, podem ser realizadas. Por fim, é interessante notar a necessidade de pesquisas por *interfaces* que suportem a realização destas idéias.

4.8 - Exibição Visual de Dados Estatísticos

O poder da estatística é inegável quando um problema específico pode ser representado através de números, especialmente quando o volume destes dados é alto. Os dados armazenados nos bancos de dados são processados por métodos estatísticos que, por sua vez, fornecem resultados que assumem a característica das informações alimentadas. Por isso, pode-se afirmar que a estatística é capaz de representar sinteticamente um conjunto de dados, como se deseja.

Tendo este conhecimento em mãos, não é difícil imaginar a utilização conjunta da estatística com as técnicas de visualização. A primeira representa os dados resumidamente, enquanto que a segunda representa os dados visualmente. São, pois, duas formas de representação de informações, uma se valendo da capacidade de síntese matemática e a outra se valendo do sistema visual humano, mas ambas mantendo a estrutura original da informação, estrutura que, primordialmente, a designa. Portanto, sobre a estrutura dos dados em análise, a visualização dos dados estatísticos se torna possível como um instrumento que envolve síntese e clareza coligados de maneira que a informação mais representativa seja explicitada.

Figura 17 - Exemplo de exibição de dados estatísticos sobre cena de visualização. Na imagem de uma cena da técnica de Coordenadas Estelares temos as médias em verde, os valores de moda em rosa, os desvios padrões em azul claro e as medianas em azul escuro. É mostrado também o menu de interação oferecido ao usuário.

Desenvolvendo este raciocínio no trabalho aqui descrito, o sistema de visualização alcançado desfruta de recursos de exibição dos principais constituintes da ciência estatística: média, mediana, desvio-padrão e moda. Na figura 17 é possível observar a exibição destes dados sobre a técnica de Coordenadas Estelares, sendo que todas as demais técnicas implementadas também receberam estas funcionalidades. Todos podendo ser utilizados concomitantemente sobre o conjunto de dados completo ou sobre porções dinamicamente selecionadas através dos recursos já mencionados.

A conclusão dos cálculos estatísticos é mostrada na cena de visualização como um elemento de informação em destaque capaz de sumarizar, caracterizar, particionar, ou diferenciar todo o conjunto, ou seleções de interesse. Assim, os testes realizados com a exibição dos dados estatísticos demonstraram um enorme potencial revelador capaz de impulsionar a análise dos dados, impulso especialmente observado quando realizado em união com a filtragem interativa num cenário que habilita a comparação entre diversos subconjuntos de interesse.

A utilização da estatística mostrou-se mais útil nas visualizações baseadas nas Coordenadas Paralelas e nas Coordenadas Estelares. Credita-se isto ao fato de que estas

visualizações, ao contrário dos *ScatterPlots* e da *Table Lens*, permitem a verificação das informações sem a alienação da unicidade dos objetos de dados. Ou seja, em ambas um registro do banco de dados é apresentado como um único elemento gráfico, uma linha ou um ponto, o que viabiliza o discernimento de síntese que se almeja. Em contrapartida, no *ScatterPlots* por exemplo, um dado n -dimensional é exibido na matriz principal através de n pontos disjuntos e isolados dentro de espaços de mapeamento não correlatos.

Dentre todas as técnicas de visualização, a mais engenhosa forma de agregação das informações estatísticas aconteceu nas Coordenadas Estelares, como se segue. Inicialmente, a idéia era desenhar pontos diferenciados cujos posicionamentos seriam dados considerando-se os resultados estatísticos como um registro especial do conjunto de dados; o que não se mostrou muito eficiente. Isso aconteceu, porque a natureza da técnica não favorece esta abordagem, já que, durante a exibição da cena, os dados utilizados para o cálculo das estatísticas não apresentavam qualquer inter-relação visual aparente, nem entre si e nem em comparação com os elementos gráficos utilizados para representar os valores estatísticos. Com o problema identificado, e perante a clareza de exibição estatística das Coordenadas Paralelas, uma abordagem mais interessante foi a utilização das idéias propostas pela técnica denominada Coordenadas Paralelas Circulares cuja descrição pode ser encontrada em (Hoffman and Grinstein 2002). Esta abordagem permite a apresentação de uma linha composta por várias semi-retas sobre um espaço bidimensional onde os eixos se arranjam da mesma forma que nas Coordenadas Estelares. A apresentação estatística aconteceu, portanto, através de um polígono cujo centro é o ponto comum a partir do qual os eixos da cena emanam. Os resultados foram encorajadores, como se observa na figura 17.

Considera-se a exibição de dados estatísticos um grande impulso às técnicas de Visualização de Informações, como se pôde observar nos testes realizados. Afirmar-se ainda que cálculos estatísticos mais sofisticados podem ser exibidos sobre as cenas de visualização, e que a implementação aqui descrita é apenas o princípio de um paradigma extremamente promissor.

4.9 - Características do Projeto de Software

Como opção de desenvolvimento, as técnicas de visualização foram implementadas como componentes de software de forma que não se tornassem específicas do sistema sendo aperfeiçoado. Esta preocupação é bastante válida devido ao fato de que os dados a serem visualizados podem originar-se de inúmeras fontes; desde o resultado do processamento de um software, um arquivo texto, ou uma conexão com uma base de dados. Assim, as *interfaces* dos componentes desenvolvidos os habilitam, pois, a ser parte integrante de sistemas de naturezas produtoras ou consumidoras de dados, podendo serem utilizados para enaltecer projetos através das facilidades do paradigma de componentes de software. A seguir é feita uma breve revisão da metodologia de desenvolvimento utilizada visando justificar sua aplicação.

Um componente de software é uma funcionalidade cuja implementação pode acontecer de maneira independente, para que seja reutilizada em um amplo escopo de aplicações (Goulão and Abreu 2002). Os componentes de software podem ser encarados como caixas pretas com *interfaces* bem definidas para utilização como constituintes de arquiteturas de projetos de software. O desenvolvimento de aplicações baseadas em componentes de software tem tido cada vez mais participação na elaboração de sistemas, pois se espera que os componentes reduzam o custo e o tempo de chegada ao mercado dos sistemas dos quais façam parte, ao mesmo tempo em que aumentem sua qualidade. A seguir apresentam-se as principais características dos componentes de software (Cai, Lyu et al. 2000):

- São partes independentes e substituíveis de sistemas que satisfazem determinadas funções;
- Os componentes funcionam dentro do contexto de uma arquitetura bem definida;
- Um componente se comunica com outros componentes através de suas *interfaces*.

Ainda, além da redução de custos e do tempo de desenvolvimento, a utilização de componentes de software é capaz de aperfeiçoar a manutenibilidade, a confiabilidade e a qualidade global dos sistemas de software (Pour, Griss et al. 1999).

Com estes objetivos em mente prosseguiu-se com a implementação das novas funcionalidades do FastMapDB, cujo desenvolvimento foi realizado com auxílio da IDE (*Interface Development Enviroment*) Borland na linguagem C++. Outra preocupação do projeto

foi o desenvolvimento orientado a objetos, cujas vantagens são bem conhecidas. Desta forma, a complexidade, inerente ao software que se objetivava, foi diluída entre entidades computacionais bem definidas, com um código mais claro e organizado.

Foi utilizada como plataforma gráfica do projeto a *Open Graphics Library* (OpenGL), um padrão aberto criado e gerido pela empresa Silicon Graphics que visa prover a desenvolvedores um ambiente de aplicação gráfica independente de plataforma, e com desempenho superior. Estes atributos decorrem do fato de que o OpenGL é disponibilizado diretamente em hardware por muitos fabricantes de dispositivos gráficos. Dessa forma, as chamadas à biblioteca são interceptadas e direcionadas ao hardware de vídeo que as converte em gráficos no dispositivo de exibição. As principais vantagens são:

- Independência de plataforma operacional;
- Desempenho superior;
- Maior portabilidade.

Portanto, o resultado final apresenta-se na forma de um módulo de software que pode ser acoplado a sistemas de qualquer natureza, ao mesmo tempo em que faz uso do potencial provido pela biblioteca OpenGL que, em determinadas configurações de hardware, é capaz de apresentar um desempenho superior ao tradicional desenvolvimento sobre sistemas de janelas embutidos nos sistemas operacionais.

4.10 - Arquitetura de Desenvolvimento

Uma das principais preocupações durante a concepção do projeto de visualização foi a criação de um código reaproveitável, que possibilitasse a implementação de outras técnicas de visualização com o mínimo de esforço, consolidando, desta forma, um Arcabouço de Visualização de Informações no âmbito de desenvolvimento de softwares. Nesse sentido, projetou-se um software baseado em camadas, sendo que nas camadas mais baixas a generalização dos algoritmos foi o princípio que orientou o trabalho.

Nas camadas inferiores (ver figura 18) ficam abstraídos o suporte à biblioteca OpenGL e aos dados pré-processados (lidos da base, normalizados, analisados e armazenados em memória). A classe *OpenGLPanel*, base da classe *VisualizationPanel*, é desenvolvida sobre o OpenGL e descende da classe *Panel* fornecida pela Borland junto com a IDE C++ Builder. Esta

classe implementa as operações necessárias para que o OpenGL seja utilizado diretamente sobre o componente *Panel*, que é muito utilizado na construção de *interfaces* gráficas e que, como classe base, possibilita que as técnicas de visualização construídas também possuam as funcionalidades de um componente de software. A classe *OpenGLPanel* também abstrai as peculiaridades de utilização de fontes sobre imagens geradas com o OpenGL, proporcionando uma *interface* de alto nível através da qual a exibição de caracteres formatados é bastante simplificada.

Figura 18 - A arquitetura dos componentes desenvolvidos. Os componentes de hardware são apresentados em cinza. Os componentes de software de terceiros são apresentados em amarelo. Em azul estão indicados os softwares desenvolvidos, ou que requerem implementação para que uma nova técnica seja desenvolvida.

Sobre o *OpenGLPanel* está a classe abstrata *VisualizationPanel*, uma especialização do *OpenGLPanel* que contém as funcionalidades comuns a todas as técnicas de visualização: funções básicas de *interface* como menus e processamento de eventos; métodos considerados necessários e declarados como abstratos obrigam, ao mesmo tempo em que guiam, o desenvolvedor a implementar as funcionalidades essenciais para a construção de técnicas visuais; propriedades que propiciam o *Link & Brush* e o cálculo estatístico; além de operações de instanciação e limpeza de memória. A classe *VisualizationPanel* é quem permite que o desenvolvimento de novas técnicas visuais ocorra de maneira rápida e simplificada; abstrata, sua utilização só é possível ao

implementarem-se funções essenciais como, por exemplo, a função que determina qual dimensão está sendo apontada pelo *mouse*, específica para cada técnica, que deve ser fornecida antes que se possa utilizar a classe.

Suportando a classe *VisualizationPanel* tem-se, como suas propriedades, as classes *PreProcessedData*, *Statistics* e *SelectionsController*. Estas classes determinam um tratamento comum dos dados em todas as técnicas visuais possibilitando a integração das mesmas, além de economia de memória. A classe *PreProcessedData* provê acesso aos dados sendo visualizados, bem como às informações de frequência e relevância utilizadas nas funcionalidades já mencionadas em seções anteriores. Esta classe guarda os dados em sua forma normalizada, o que possibilita a apresentação dos mesmos por diversas técnicas visuais e, ao mesmo tempo em que é acessada diretamente pela classe *VisualizationPanel*, também dá suporte às classes *Statistics* e *SelectionsController* que realizam operações sobre os dados.

Como se sabe, cada uma das técnicas de visualização possui uma forma peculiar de apresentação, determinando que as seleções sobre os dados e a apresentação da estatística calculada sobre eles ocorra de maneira única para cada contexto. No entanto, as classes *Statistics* e *SelectionsController* são únicas e acessadas como propriedades da classe *VisualizationPanel*, sem qualquer especialização, como se espera de uma classe base. Para que esta arquitetura fosse possível, as classes *Statistics* e *SelectionsController* foram organizadas em dois aspectos. O primeiro aspecto destas classes implementa o processamento a partir de um conjunto mínimo de informações, atualizando as informações estatísticas relevantes ou redefinindo as seleções sobre os conjuntos de dados, respectivamente. O segundo aspecto se restringe a realizar chamadas a métodos de desenho fornecidos às classes através de suas *interfaces*.

As classes *Statistics* e *SelectionsController* recebem dados fornecidos por uma *interface* bem definida acessada pela classe *VisualizationPanel* que, desta forma, utiliza-se do processamento previsto pelas duas classes. Da mesma maneira, os métodos de desenho das estatísticas e das seleções de dados são fornecidos via propriedades especiais que permitem a resolução dinâmica de chamadas funções, sendo que cada técnica de visualização implementada fica responsável por definir como os dados referentes a seleções e a estatísticas são mais bem apresentados visualmente.

A estrutura descrita culminou em um código flexível e reutilizável, a partir do qual novas técnicas de visualização gozando das funcionalidades descritas podem ser construídas. Além

disso, a utilização deste esquema resulta na possibilidade de integração dos componentes que vierem a ser criados com os já existentes. A sequência de tarefas necessárias para a implementação de uma nova técnica é descrita a seguir:

- 1) Criar uma nova classe descendendo de *TVisualizationPanel* e implementar seus métodos abstratos segundo a natureza da técnica que será criada;
- 2) Fazer a implementação mínima necessária para que as cenas de visualização possam ser geradas com base nos dados pré-processados. A partir deste código inicial será possível ter-se uma visão mais clara dos detalhes de implementação e das características do que se quer criar;
- 3) Criar uma classe derivando de *TVisualizationMethods* e implementar as funções de desenho estatístico e de seleção de acordo com as cenas que se deseja gerar. A classe de visualização terá uma propriedade *TVisualizationMethods*, assim o código de desenho que será utilizado por outras classes fica separado, facilitando o desenvolvimento do trabalho;
- 4) Implementar as funcionalidades mais refinadas como o *Frequency Plot* e o *Relevance Plot*;
- 5) Apurar as cenas através das possibilidades do OpenGL.

O desenvolvimento do sistema deu-se sobre uma arquitetura em camadas com o código seguindo o paradigma orientado a objetos, preservando-se as *interfaces*, pode-se expandí-lo de maneira natural.

4.11 - O Pipeline de Visualização

Com a integração dos componentes que implementam as técnicas de visualização com a ferramenta FastMapDB, foi alcançado um novo exemplo que serve como modelo, ou seja, um novo paradigma de visualização de dados. Constituiu-se assim um *Pipeline* de visualização. Como contribuição deste trabalho, este novo modo de exploração visual de dados pode ser traduzido como um canal por onde a informação flui, passando por filtros cognitivos gerados dinamicamente via intervenção direta do usuário, alcançando progressivamente maiores graus de refinamento segundo os valores a que a informação está sujeita. No trabalho aqui descrito, o fluxo da informação passa por dois níveis:

- 1) O poder de investigação da tradicional visualização tridimensional no plano cartesiano foi explorada no FastMapDB, o que se traduz na análise visual prévia via a projeção tridimensional dos dados, na qual os detalhes são omitidos;
- 2) A reutilização dos produtos da etapa anterior através de técnicas de introspecção visual providas da vanguarda da ciência de Infovis, ou seja, a análise visual multimodal de porções identificadas como interessantes, na qual a informação pode ser minuciosamente explorada por diversos métodos.

Apesar de passível de descrição, as reais propriedades do sistema alcançado, acredita-se, só podem ser percebidas via sua utilização. Ademais, sua natureza gráfica interativa gera resultados segundo o que se observa em inúmeras configurações distintas da cena de visualização e, a ilustração deste processo esbarra na natureza antagônica da apresentação do texto, estática e com limitação espacial. Portanto, procura-se ao máximo transmitir o conteúdo da matéria defendida através de *screenshots* de exemplos significativos. Como se segue:

Como exemplo será utilizada uma base de dados com indicadores habitacionais de 327 cidades norte-americanas. Cada registro, representando uma cidade, possui 9 atributos: região climática, índice habitacional de casas, saúde, criminalidade, transporte, educação, artes, recreação e economia. Com exceção do primeiro atributo, os demais podem ser interpretados através de uma análise comparativa de grandezas, isto é, as cidades podem ser comparadas com base na magnitude relativa de seus atributos. A visualização desta base no FastMapDB, figura 19(a), revela que existem duas caracterizações das cidades que compõem os dados: há um aglomerado bem denso e definido e outro bastante esparsa e distribuído. Não há um atributo classificador.

Ainda na figura 19 podem ser observadas duas seleções realizadas sobre o cubo interativo da ferramenta e suas respectivas visualizações apresentadas através da técnica das Coordenadas Estelares: na figura 19(b) o aglomerado menor e mais bem definido, e na figura 19(c) o maior e mais esparsa. A figura 19(b) revela que o primeiro agrupamento é composto de cidades com mais educação, arte e índice habitacional de casas, e o segundo agrupamento, figura 19(c), composto de cidades com mais recreação, economia e transporte. Portanto, pode-se inferir que os grupos se referem, respectivamente, a cidades pequenas e médias no menor aglomerado, e a cidades grandes no maior. E de fato, entre as cidades compreendidas pela maior partição da cena, temos Nova York, São Francisco, Detroit e Los Angeles, conhecidos grandes centros dos EUA.

Através deste exemplo procurou-se evidenciar o potencial exploratório proporcionado pela refinação sucessiva dos dados através de meios diversificados de análise. Isso ocorre apesar do fato da implementação alcançada determinar um fluxo de informações limitado a apenas dois níveis. Observe, que a idéia proposta pode ser estendida a uma quantidade maior de estágios de análise. Por exemplo, um terceiro nível poderia compreender técnicas de MD algorítmicas aplicadas a seleções realizadas sobre partições dos dados escolhidas visualmente. O princípio é o refinamento sucessivo dos dados por meios diversificados de análise que, se propriamente esquematizados, podem ser de grande valia; acredita-se que a proposta possa gerar uma nova linha de trabalhos seguindo este fundamento.

Figura 19 - Exemplo de utilização do *Pipeline* de visualização. Em (a) a visualização global dos dados referentes às cidades; os detalhes verdes indicam os agrupamentos selecionados para visualização multimodal. Em (b) temos a visualização, por Coordenadas Estelares, do agrupamento menor; em (c) a visualização do maior.

4.12 - Considerações Finais

Neste capítulo foram mostrados os detalhes das realizações e as contribuições conseguidas durante a execução do projeto, bem como os caminhos trilhados para alcançar tais objetivos. Em conjunto com a exposição dos tópicos foram propostas breves discussões que levantaram críticas, sugestões e avaliações sobre os aspectos merecedores de atenção.

Foi apresentada a ferramenta FastMapDB, marco inicial do projeto, juntamente com uma breve discussão de seu princípio de funcionamento, suas características e operação. As contribuições diretas à ferramenta foram vistas na sequência, bem como os resultados imediatos de sua utilização. Foi feita, então, a revisão crítica e comparativa das técnicas de Visualização de Informações utilizadas no projeto e como se deu a união destas técnicas para que se tivesse um ambiente altamente integrado através do princípio de *Link & Brush*. As seções seguintes do capítulo apresentaram as definições e propostas utilizadas para as implementações da exibição visual de frequências e da exibição de dados por relevância, juntamente com as respectivas discussões sobre utilização e possibilidades de extensão das propostas. Foi também detalhada a utilização de dados estatísticos para fomentar as cenas geradas segundo as técnicas de visualização, a exposição foi feita focando-se na implementação realizada no projeto.

As principais características do software desenvolvido bem como sua metodologia de desenvolvimento, utilizando o conceito de componentes de software, orientação a objetos e suporte da biblioteca gráfica OpenGL foram apresentadas. Em seguida foi colocada a arquitetura do Arcabouço determinado pelo trabalho finalizado.

Finalizando, este capítulo discorreu-se a respeito do *Pipeline* de visualização, o qual foi a motivação inicial deste projeto. Procurou-se descrever a idéia desenvolvida mostrando-a como um novo paradigma para a melhor compreensão de conjuntos de dados.

Capítulo 5 - Conclusões e Linhas de Futuras Pesquisas

5.1 - Considerações Finais

No trabalho realizado no decorrer do Programa de Mestrado foram revisados conceitos e tópicos relacionados com o tema de Visualização de Informações em bases de dados. Portanto, visando a maestria nestes rumos da ciência, foram compilados temas correlatos que perfazem o contexto através do qual a Visualização de Informações pôde ser concretizada como especialização do curso. Desta forma, bancos de dados, Descoberta de Conhecimento em Bases de Dados e Mineração de Dados precederam o assunto principal, a Visualização de Informações, numa abordagem teórica que precedeu a apresentação dos resultados práticos concretizados no decorrer das atividades do projeto.

Na apresentação dos resultados (capítulo 4), parte principal da dissertação, foram descritas as contribuições do trabalho concluído. Tais contribuições são frutos das atividades de implementação que consumiram a maior parte do tempo reservado ao projeto. Na implementação foram aplicados conceitos novos que propiciam o desenvolvimento de um software mais robusto e portátil, além de contribuírem para a formação do aluno. Princípios de orientação a objetos e componentes de software construídos sobre a plataforma OpenGL resultaram em um trabalho mais elaborado e reaproveitável, capaz de servir como exemplo para a construção de outros sistemas que agreguem a Visualização de Informações.

As realizações do projeto foram: as propostas denominadas de *Frequency Plot* e *Relevance Plot*, o *Pipeline* de visualização, as métricas de identificação de aglomerados adicionadas ao FastMapDB e, por fim, a exibição visual de estatísticas sobre as imagens dos dados, tudo na forma de software reutilizável. Desta forma o Arcabouço de visualização não se resumiu à simples implementação de técnicas visuais já bastante conhecidas e de domínio público. Os componentes de software que foram criados carregam possibilidades de novas experimentações

no campo da análise de dados, seja através da estatística, da exibição de frequências, ou mesmo da exploração por relevância. A integração de todo este potencial é um ponto a mais a ser considerado, e ao analista é oferecida a possibilidade de navegar entre diversos ângulos das imagens dos dados.

Portanto, segundo o que se concluiu com base nas pesquisas, a orientação do projeto foi de encontro ao maior problema das atuais propostas de análise visual: a ausência de interatividade das imagens dos dados. É o que se viu em diversos pacotes de visualização disponíveis na *web*, nos quais a falta de recursos de exibição não convencional dos dados era marcante. E, mesmo quando meios de interação com as imagens foram proporcionados, muitas das vezes estes eram ineficientes, de difícil utilização, especialmente devido ao alto poder computacional demandado pelos softwares. Portanto, através dos resultados encorajadores alcançados neste trabalho, acredita-se que a real utilização das técnicas de visualização de dados só alcançará sua plenitude em um sistema que proporcione concretas ferramentas de manipulação dos elementos gráficos, propiciando ao usuário a sensação tátil objetiva dos elementos de informação representados. Assim, aliando-se meios analíticos numéricos e automatizados a formas eficientes de apresentação de seus resultados, os projetos de análise visual encontrarão um forte paradigma de arquitetura de software que trará a ciência de Infovis ao âmbito do aproveitamento dos dados mesmo em ambientes não corporativos. Da mesma forma como se observa em técnicas convencionais de projeção 2D, presentes em qualquer programa de planilha de cálculos.

5.2 - Sugestões de Futuras Pesquisas

Como continuidade imediata do presente trabalho, a seguir são colocadas três sugestões:

- Integração com técnicas de Mineração de Dados tanto para exploração de dados como para apresentação dos resultados da mineração, aumentando as possibilidades analíticas;
- Avaliação qualitativa mais aprofundada das técnicas de Visualização de Informações visando a utilização de métricas existentes ou novas para avaliar a utilização dos conceitos propostos;
- Tratamento do problema de escalabilidade das técnicas de Visualização de Informações através de processos de amostragem para conjuntos de dados muito grandes, ampliando o escopo de utilização dos processos visuais de exploração para conjuntos extremamente grandes.

Referências Bibliográficas

- Ahlberg, C. and B. Shneiderman (1994). Visual Information Seeking: Tight coupling of Dynamic Query Filters with Starfield Displays. Proc. Human Factors in Computing Systems CHI '94.
- Albuz, E., E. Kocalar, et al. (2001). Scalable Color Image Indexing and Retrieval Using Vector Wavelets. IEEE Transactions on Knowledge and Data Engineering, ACM Press.
- Alpern, B. and L. Carter (1991). The hyperbox. Proceedings of the IEEE Conference on Visualization, San Diego, CA, USA.
- Ankerst, M., C. Elsen, et al. (1999). Visual Classification: An Interactive Approach to Decision Tree Construction. ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, San Diego, CA, USA, ACM Press.
- Ankerst, M., M. Ester, et al. (2000). Towards an Effective Cooperation of the User and the Computer for Classification. ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining (KDD'2000), Boston, MA, ACM Press.
- Baeza-Yates, R. and B. Ribeiro-Neto (1999). Modern Information Retrieval. Wokingham, UK, Addison-Wesley.
- Baeza-Yates, R. and B. A. Ribeiro-Neto (1999). Modern Information Retrieval. Wokingham, UK, Addison-Wesley.
- Berchtold, S., C. Böhm, et al. (1997). A Cost Model For Nearest Neighbor Search in High-Dimensional Data Space. ACM Symp. on Principles of Database Systems (PODS), Tucson, AZ, ACM Press.
- Berchtold, S., C. Böhm, et al. (1998). The Pyramid-Tree: Breaking the Curse of Dimensionality. ACM Int'l Conference on Data Management (SIGMOD), Seattle, WA.
- Beyer, K., J. Godstein, et al. (1999). When is "Nearest Neighbor" Meaningful? 7th International Conference (ICDT '99), Jerusalem, Israel, Springer.
- Bier, E. A., M. C. Stone, et al. (1993). Toolglass and Magic Lenses: The See-Through Interface. SIGGRAPH '93.
- Böhm, C. and H.-P. Kriegel (2000). Dynamically Optimizing High-Dimensional Index Structures. Intl. Conf. on Extending Database Technology, Kontanz, Germany, Springer Verlag.

- Bueno, J. M. (2002). Suporte à Recuperação de Imagens Médicas baseada em Conteúdo através de Histogramas Métricos. Departamento de Ciências de Computação. São Carlos, SP, Universidade de São Paulo: 146.
- Cabena, P., P. Hadjinian, et al. (1998). Discovering Data Mining: From Concept to Implementation. Englewood Cliffs, NJ, Prentice Hall.
- Cai, X., M. R. Lyu, et al. (2000). Component-Based Software Engineering: Technologies, Development Frameworks, and Quality Assurance Schemes. Proceedings of the Seventh Asia-Pacific Software Engineering Conference (APSEC.00).
- Card, S. K., J. D. Mackinlay, et al. (1999). Using Vision to Think. San Francisco, CA, Morgan Kaufmann Publishers.
- Castañón, C. A. B. (2003). Recuperação de Imagens por Conteúdo através de Análise Multi-escala por Wavelets. Ciências de Computação e Estatística. São Carlos, SP, Universidade de São Paulo: 95.
- Chambers, J. M., W. S. Cleveland, et al. (1983). Graphical Methods for Data Analysis. Belmont, CA, Wadsworth International Group.
- Chen, M.-S., J. Han, et al. (1996). "Data Mining: An Overview from a Database Perspective." IEEE Transactions on Knowledge and Data Engineering 8(6): 866-883.
- Chernoff, H. (1973). "The Use of Faces to Represent Points in k-Dimensional Space Graphically." Journal of the American Statistical Association 68(342): 361 - 368.
- Chi, E. H. (2002). A Framework for Visualizing Information. Palo Alto, CA, Kluwer Academic Publishers.
- Clifton, C. and D. Marks (1996). Security and Privacy Implications of Data Mining. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery.
- Duke, D. (2001). Modular techniques in information visualization. Australian symposium on Information visualisation, Sydney, Australia, Australian Computer Society, Inc.
- Faloutsos, C. and K. Lin (1995). FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. ACM Int'l Conference on Data Management (SIGMOD), Zurich, Switzerland, Morgan Kaufmann.
- Fayyad, U. (1998). "Mining Databases: Towards Algorithms for Knowledge Discovery." Bullettin of Tech. Committee on Data Engineering 21(1): 29-48.

- Fayyad, U., G. Piatetsky-Shapiro, et al. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM. **39**: 27 - 34.
- Fayyad, U. M., G. Piatetsky-Shapiro, et al. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, USA, AAAI Press.
- Foster, M. and A. G. Gee (2002). The Data Visualization Environment. Information Visualization in Data Mining and Knowledge Discovery. U. Fayyad, G. G. Grinstein and A. Wierse, Morgan Kaufmann Publishers: 87-93.
- Frawley, W. J., G. Piatetsky-Shapiro, et al. (1991). Knowledge Discovery in Databases: An Overview. P.-S. e. Frawley, AAAI/MIT Press: 1-27.
- Fua, Y.-H., M. O. Ward, et al. (1999). "Hierarchical Parallel Coordinates for Exploration of Large Datasets." Proc. IEEE Visualization'99.
- Ganesh, M., E. Han, et al. (1996). Visual Data Mining: Framework and Algorithm Development. Minneapolis, MN, University of Minnesota.
- Goulão, M. and F. B. Abreu (2002). The Quest for Software Components Quality. Proceedings of the 26 th Annual International Computer Software and Applications Conference (COMPSAC'02), Oxford, England.
- Grinstein, G. G., M. Trutschl, et al. (2001). High-Dimensional Visualizations. Knowledge Discovery and Data Mining - Workshop on Visual Data Mining, San Francisco, California, USA.
- Grinstein, G. G. and M. O. Ward (2002). Introduction to Data Visualization. Information Visualization in Data Mining and Knowledge Discovery. U. Fayyad, G. G. Grinstein and A. Wierse, Morgan Kaufmann Publishers: 21-45.
- Han, J. and M. Kamber (2001). Data Mining - Concepts and Techniques. New York, Morgan Kaufmann Publishers.
- Hinneburg, A., D. A. Keim, et al. (1999). "HD-Eye: Visual Mining of High-Dimensional Data." IEEE Computer Graphics and Applications **19**(5): 22-31.
- Hirji, K. K. (2001). Exploring Data Mining Implementation. Communications of the ACM. **44**: 87 - 93.

- Ho, T. B. and T. D. Nguyen (2001). Visualization Support for User-Centered Model Selection in Knowledge Discovery in Databases. 13th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'01), Dallas, Texas.
- Hoffman, P. E. and G. G. Grinstein (2002). A Survey Of Visualizations for High-Dimensional Data Mining. Information Visualization in Data Mining and Knowledge Discovery. U. Fayyad, G. G. Grinstein and A. Wierse, Morgan Kaufmann Publishers: 47-82.
- Hoffman, P. E., G. G. Grinstein, et al. (1997). DNA visual and analytic data mining. IEEE Visualization, Phoenix, Arizona, USA.
- Inselberg, A. (1985). "The Plane with Parallel Coordinates." The Visual Computer 1(Special Issue on Computational Geometry): 69-91.
- Inselberg, A. and B. Dimsdale (1990). Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry. IEEE Visualization, IEEE Computer Press.
- Kandogan, E. (2000). Star Coordinates: A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions. IEEE Symposium on Information Visualization 2000, Salt Lake City, Utah.
- Kandogan, E. (2001). Visualizing Multi-dimensional Clusters, Trends, and Outliers using Star Coordinates. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA.
- Keim, D. and H. P. Kriegel (1995). Issues in Visualizing Large Databases. Proceedings of the International Conference on Visual Database Systems, Lausanne, Switzerland.
- Keim, D., J. P. Lee, et al. (1995). Database Issues for Data Visualization: Supporting Interactive Database Exploration. Workshop on Database Issues for Data Visualization - IEEE Visualization '95, Atlanta, Georgia, USA, Springer.
- Keim, D. A. (1997). Visual Database Exploration Techniques. KDD'97 Int. Conf. on Knowledge Discovery and Data Mining, Newport Beach, CA.
- Keim, D. A. (2000). Designing Pixel-Oriented Visualization Techniques: Theory and Applications. IEEE Trans. on Visualisation and Computer Graphics, IEEE Computer Society. **6**: 59-78.
- Keim, D. A. (2002). "Information Visualization and Visual Data Mining." IEEE Transactions on Visualization and Computer Graphics **8**(1): 1-8.

- Keim, D. A., M. Ankerst, et al. (1995). Recursive Pattern: A technique for Visualizing Very Large Amounts of Data. Proceedings of the 6th IEEE Visualization Conference, Atlanta, GA.
- Keim, D. A., M. C. Hao, et al. (2001). "Pixel bar charts: a visualization technique for very large multi-attributes data sets." Information Visualization Journal.
- Keim, D. A. and H.-P. Kriegel (1994). "VisDB: Database Exploration Using Multidimensional Visualization." IEEE Computer Graphics and Applications **14**(5): 16-19.
- Keim, D. A. and H.-P. Kriegel (1996). "Visualization Techniques for Mining Large Databases: A Comparison." IEEE Transactions in Knowledge and Data Engineering **8**(6): 923-938.
- LeBlanc, J., M. O. Ward, et al. (1990). Exploring N-Dimensional Databases. Proc. IEEE Visualization'90, San Francisco, CA.
- Lee, T. S. (1996). Image Representation Using 2D Gabor Wavelets. IEEE Transactions on Pattern Analysis and Machine Intelligence, ACM Press.
- Leung, Y. and M. Apperley (1994). A Review and Taxonomy of Distortion-Oriented Presentation Techniques. Human Factors in Computing Systems CHI '94.
- Lutu, P. E. N. (2002). An Integrated Approach for Scaling up Classification and Prediction Algorithms for Data Mining. Proceedings of the 2002 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology, Port Elizabeth, South Africa, South African Institute for Computer Scientists and Information Technologists.
- Lux, M. (1998). Level of Data - A Concept for Knowledge Discovery in Information Spaces. International Conference on Information Visualisation, London, England.
- Mackinlay, J. D., G. G. Robertson, et al. (1991). The perspective wall: detail and context smoothly integrated. Conference on Human Factors and Computing Systems, New Orleans, Louisiana, USA.
- Martin, A. R. and M. O. Ward (1995). High Dimensional Brushing for Interactive Exploration of Multivariate Data. 6th IEEE Visualization Conference, Atlanta, Georgia, USA.
- Miller, J. J. and E. J. Wegman (1991). Construction of line densities for parallel coordinate plots. Computing and Graphics in Statistics, Springer-Verlag: New York, Buja, A.
- Tukey, P.
- Munzner, T. (2002). Guest editor's introduction - information visualization. IEEE Computer Graphics and Applications. **22**: 20-21.

- Oliveira, M. C. F. and H. Levkowitz (2002). From Visual Data Exploration to Visual Data Mining: A Survey. IEEE Transactions on Visualization and Computer Graphics.
- Pickett, R. M. and G. G. Grinstein (1988). Iconographic Displays for Visualizing Multidimensional Data. IEEE Conf. on Systems, Man and Cybernetics, Piscataway, NJ, IEEE Press.
- Pour, G., M. Griss, et al. (1999). Making the Transition to Component-Based Enterprise Software Development: Overcoming the Obstacles - Patterns for Success. Proceedings of Technology of Object-Oriented Languages and Systems, China.
- Rao, R. and S. K. Card (1994). The Table Lens: Merging Graphical and Symbolic Representation in an Interactive Focus+Context Visualization for Tabular Information. Proc. Human Factors in Computing Systems.
- Rui, Y., T. S. Huang, et al. (1999). "Image retrieval: current techniques, promising directions and open issues." Journal of Visual Communication and Image Representation **10**: 1-23.
- Rundensteiner, A., M. O. Ward, et al. (2002). Xmdv Tool: Visual Interactive Data Exploration and Trend Discovery of High Dimensional Data Sets. Proceedings of the 2002 ACM SIGMOD international conference on Management of data, Madison, Wisconsin, USA, ACM Press.
- Sammon, J. W., Jr. (1969). A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers. **C-18(5)**: 401-409.
- Santos, R. F., Filho, A. J. M. Traina, et al. (2001). Similarity Search without Tears: The OMNI Family of All-purpose Access Methods. Intl. Conf. on Data Engineering (ICDE), Heidelberg, Germany, IEEE Computer Society.
- Sarkar, M. and M. Brown (1994). "Graphical Fisheye Views." Communications of the ACM **37(12)**: 73-84.
- Schneiderman, B. (1996). The Eyes Have it: A Task by Data Type Taxonomy of Information Visualizations. IEEE Symposium on Visual Languages, IEEE Computer Society.
- Shneiderman, B. (1984). Response time and display rate in human performance with computers. ACM Computing Surveys (CSUR). **16**: 265-285.
- Shneiderman, B. (1992). Tree Visualization with Treemaps: A 2D Space-Filling Approach. ACM Trans. Graphics. **11**: 92-99.

- Siirtola, H. (2000). Direct Manipulation of Parallel Coordinates. International Conference on Information Visualization.
- Traina, A. J. M., T. Caetano, Jr, et al. (2001). Visualização de Dados em Sisemas de Bases de Dados Relacionais. XVI Simpósio Brasileiro de Banco de Dados, Rio de Janeiro, SBC.
- Traina, A. J. M. and C. F. Paniago (-sFile1994). Compression of Images: A Modular System with Multiformat Management. Proceedings of the VII Brazilian Conference on Computer Graphics and Image Processing.
- Traina, C., Jr, A. J. M. Traina, et al. (1999). FastMapDB User's Manual. Carnegie Mellon University - School of Computer Science: 8p.
- van Wijk, J. J. and R. van Liere (1993). HyperSlice. Proceedings of the IEEE Conference on Visualization, San Jose, CA, USA, IEEE Computer Society Press.
- Vieira, M. T. P., M. Biajiz, et al. (1999). Metadata for Content-Based Search on an MHEG-5. Proceedings of the Third IEEE META-DATA Conference, Bethesda, Maryland, USA.
- Walter, J. and H. Britter (2002). On Interactive Visualization of high-dimensional Data using the Hyperbolic Plane. 8th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining, Alberta, CA.
- Ward, M. O. (1994). XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data. Proceedings of IEEE Conference on Visualization.
- Ward, M. O. (1997). Creating and Manipulating n-dimensional brushes. American Statistics Association - Section on Statistical Graphics.
- Wegman, E. J. and Q. Luo (1997). "High Dimensional Clustering Using Parallel Coordinates and the Grand Tour." Computing Science and Statistics **28**: 352–360.
- Wong, P. C. (1999). "Visual Data Mining." IEEE Computer Graphics and Applications **19**(5): 20-21.
- Wong, P. C. and R. D. Bergeron (1995). Multiresolution multidimensional wavelet brushing. Proceedings of IEEE Visualization, Los Alamitos, CA, IEEE Computer Society Press.