Referência: Visualização como Suporte à Exploração de uma Base de Dados Pluviométricos, dissertação de mestrado de autoria de Vinícius M. A.Branco e orientação Profa. Dra. Maria Cristina F. de Oliveira,Capítulo 2, fev/2003, ©ICMC-USP.

SCE 236 Visualização Computacional

Visualização de Informação e Mineração Visual de Dados

Mineração Visual de Dados é um tema que vem recebendo destaque no meio acadêmico, tanto pelas expectativas atribuídas aos promissores benefícios oriundos de suas funcionalidades, quanto pelas dificuldades características do processo, cujas técnicas se propõem a tratar. Unindo Análise Exploratória Visual, uma das vertentes da Visualização de Informação, e Mineração de Dados, VDM objetiva auxiliar o processo de aquisição de conhecimento utilizando-se de representações gráficas para explorar e/ou analisar grandes bases de dados. Desta forma, envolve conceitos e termos referentes a estas duas áreas.

Este capítulo apresenta os principais aspectos envolvidos com ambos os temas, Visualização de Informação e Mineração de Dados, nas *seções 2.1* e *2.2*, respectivamente. A *seção 2.1* subdivide-se em seções que descrevem dois aspectos centrais da Visualização de Informação: os conjuntos de dados e as próprias técnicas de visualização. A *seção 2.3* encerra este capítulo analisando como pesquisadores vêm empenhando esforços na tentativa de integrar os dois temas descritos nas seções anteriores.

2.1 Visualização de Informação

Representações gráficas de toda sorte têm sido usadas como instrumento de comunicação desde os primórdios da humanidade. Com o advento da ciência, as representações gráficas passam a embutir significado cada vez mais regido por convenções, a exemplo de gráficos matemáticos e cartas cartográficas. Normalmente, essas representações têm como propósito comunicar uma idéia que já existe. Todavia, tendo como propósito aproveitar as especiais características da percepção visual humana para a resolução de problemas lógicos, uma segunda abordagem possível consiste em utilizar as representações gráficas para criar ou descobrir a própria idéia.

Dentre essas duas abordagens, esta última tem sofrido de forma mais impactante as inovações geradas pela evolução dos computadores, os quais proporcionam meios cada vez mais eficientes de melhorar a geração das imagens dessas representações e de aumentar a interatividade em tempo-real, tendo, em paralelo, custo cada vez mais baixo. Esse meio permite descrições gráficas que automaticamente reúnem milhares de objetos de dados em uma figura, revelando padrões ocultos.

Entende-se por Visualização o processo de mapeamento de dados e informações em um formato gráfico, baseando-se em representações visuais e em mecanismos de interação, fazendo uso de suporte computacional e objetivando a ampliação da cognição[1]. Card et al. [Car 1999] exprimem com concisão essa idéia: "o propósito da visualização é a percepção [*insight*], não figuras"; sendo que os principais objetivos dessa percepção são a descoberta, a tomada de decisões e o entendimento.

As etapas essenciais a serem consideradas no uso de determinada técnica de Visualização, ou no desenvolvimento de novas técnicas, podem ser identificadas por meio de modelos de referência de Visualização. Uma descrição de Visualização como uma seqüência (*figura 2.1*) de mapeamentos "ajustáveis" de dados[2] para uma representação visual é dada por Card et al. [Car 1999]. Ela possibilita a interação do usuário com o espaço de informação, a fim de alcançar o que foi chamado de *cristalização do conhecimento*.

[1] Neste contexto, cognição significa aquisição ou uso de conhecimento.
[2] O modelo assume que os dados estarão na forma de tabelas de registros.

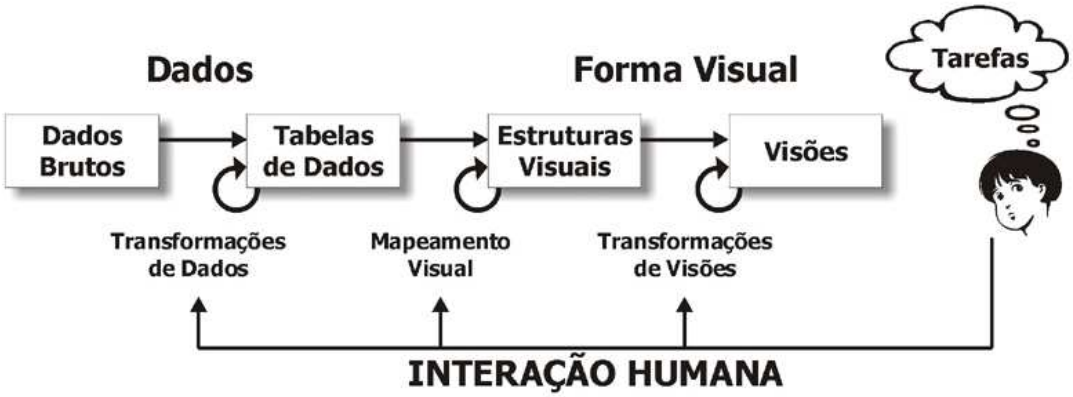

***Figura 2.1** – Modelo de Referência de Visualização de Card et al. [Car 1999]*

Muitas áreas de atuação humana estão interessadas na criação de artefatos visuais, e muitas dessas atividades foram beneficiadas com a produção de visualizações. Particularmente, a área científica vem se beneficiando do poder computacional e das visualizações sofisticadas alcançadas com ele, na chamada "Visualização Científica". Essa classe de Visualização se baseia em dados produzidos por fenômenos "naturais", do mundo f ísico, ainda que esta não tenha sido necessariamente a sua concepção original. O computador é, então, usado para tornar visível alguma propriedade de objetos de interesse. Enquanto essas visualizações podem derivar de abstrações do espaço físico, a informação é, todavia, inerentemente geométrica. Por exemplo, para representar as concentrações e a dinâmica de massas de ar na atmosfera pode-se fazer uso de abstrações, mas elas são baseadas num espaço físico.

Não obstante, há uma grande quantidade de informação não-física no mundo contemporâneo, tais como coleções de documentos e dados financeiros, que também demandam meios que proporcionem cognição. A diferença sobresselente está na total inexistência de um mapeamento espacial óbvio, resultando num problema ainda maior do que o de tornar visíveis as propriedades do objeto de interesse: o de como mapear abstrações não-espaciais em uma representação visual efetiva. O esforço de estender a visualização para a área da abstração de fenômenos não-físicos recebe o nome de "Visualização de Informação", e tem sido motivado não só pelo grande acúmulo de informações na atualidade, mas também pela complexidade intrínseca ao processo de se alcançar um mapeamento visual eficaz.

Outra razão para a distinção entre essas duas formas de Visualização, Visualização Científica e Visualização de Informação, diz respeito à própria evolução que as áreas têm apresentado [Car 1999]. Visualização de Informação nota-se originada pelo emprego de representações visuais para representar dados abstratos em diversas comunidades de pesquisa (abordada por estudos semiológicos, aplicada a dados estatísticos, etc.), e só mais tarde percebeu-se a convergência para o uso de poder computacional nas representações. Por sua vez, a Visualização Científica tem sua origem notoriamente ligada às ciências que abordam o espaço físico (física, engenharia, etc.). A distinção entre essas vertentes da Visualização é aceita por grande parte dos autores encontrados na literatura de Visualização.

Uma última observação quanto ao papel da Visualização para a obtenção de cognição, refere-se ao fato dela fornecer uma faixa maior de elementos mais facilmente distinguíveis, quando comparada aos outros sentidos de percepção humana. Embora, de forma mais ampla, a representação de dados abstratos possa ser realizada por meio de sistemas que abordem múltiplas formas de percepção, tais como *audição* e *tato*, é de senso comum que a Visualização seja apontada como um ponto inicial nessa discussão [Car 1999; Shn 1996].

2.1.1 Conjuntos de Dados

O elemento central no modelo de referência de Visualização da *figura 2.1*, é o mapeamento visual das tabelas de dados para a estrutura visual, servindo como intermediário entre as etapas que envolvem tratamento de dados e as etapas que lidam com a forma visual. As tabelas de dados baseiam-se em relações matemáticas, enquanto as estruturas visuais são baseadas nas propriedades gráficas efetivamente processadas pela visão humana. A terminologia encontrada na literatura, no que se refere a dados, não é consistente, já que vem sendo criada por muitas disciplinas (estatística, engenharia, ciências de computação, etc.) [Won 1997].

Card et. al. [Car 1999] propõem uma terminologia para dados, a fim de uniformizar o tratamento dos termos envolvidos nesse assunto, ao passo que ressaltam a tentativa de apresentar um equilíbrio entre formalidade e clareza. Segundo eles, a primeira relaciona-se com a precisão (a qual é fator crítico quando se discute o assunto "dados"), porque diferenças sutis no s dados

podem ocasionar grandes diferenças na escolha da visualização. Todavia, a clareza é importante, principalmente quando se está começando a ter conhecimento sobre as técnicas de Visualização.

2.1.1.1 Tabelas de Dados

A aplicação de relações ou conjuntos de relações estruturadas para tornar mais fácil o mapeamento para representações visuais é a estratégia mais comum para o tratamento de dados brutos (que apresentam um formato qualquer). Matematicamente, uma relação é um conjunto de tuplas:

$$\{ <valor_{ix}, valor_{iy}, ...>, <valor_{jx}, valor_{jy}, ...>, ...\}$$

A representação puramente matemática omite descrições de informação que são importantes para a Visualização. Na terminologia de Card et al., as tabelas de dados combinam relações com metadados, os quais descrevem essas relações (similar às tabelas de dados de bancos de dados relacionais):

Tabela 2.1 *– Descrição de uma tabela de dados [Car 1999]*

	Caso$_i$	**Caso$_j$**	**Caso$_k$**	...
Variável $_x$	Valor$_{ix}$	Valor$_{jx}$	Valor$_{kx}$	...
Variável $_y$	Valor$_{iy}$	Valor$_{jy}$	Valor$_{ky}$	...
...	...	...	...	...

Assim, a distribuição dos valores é feita por meio de uma matriz de casos[3] e variáveis[4], como representado na *tabela 2.1*, em que os casos são dispostos em colunas e as variáveis dispostas em linhas[5]. Normalmente, as tabelas de dados são formadas por casos independentes, ou seja, cada um deles representa uma única relação. Quando existe dependência entre casos, como nas relações pai/filho de uma árvore, eles podem ser reorganizados em um conjunto de casos independentes, de forma a obter um "*flat file*", numa técnica própria conhecida como "de-normalização" [Wit 2000]. Basicamente, esta técnica consiste na adição de um ou mais atributos

[3] É encontrado na literatura também com os nomes instâncias, registros e itens de dados.
[4] É encontrada na literatura também com os nomes atributos e dimensões, dentre outros.
[5] A escolha de qual termo recebe orientação horizontal ou vertical varia na literatura.

que registra(m) o(s) relacionamento(s) de interesse entre os casos, o que sempre pode ser realizado com qualquer conjunto (finito) de relações (finitas). Uma outra forma de organizar essas relações hierárquicas é descrevê-las por meio de arquivos estruturados e, então, visualizá-las com técnicas específicas para esse tipo de dados (hierárquicos).

Wong e Bergeron [Won 1997] utilizam os termos "multidimensional" e "multi variada" para referenciar a dimensionalidade das variáveis independentes e das variáveis dependentes, respectivamente, de um conjunto de dados. Todavia, na prática, esses termos são usados indistintamente para indicar os atributos (dimensões) associados a cada item de dados.

Embora o volume de casos normalmente seja muito maior do que o número de variáveis, uma característica cada vez mais comum nas bases de dados é a presença de um grande número de variáveis, ou dimensões (alta dimensionalidade). Por exemplo, para empresas terem Vantagem Competitiva sobre seus concorrentes, um fator diferencial é a posse de informação valiosa sobre clientes; na tentativa de obter essa informação, o que normalmente ocorre é a observação de muitos parâmetros no processo de coleta dos dados [Kei 2001]. Conseqüentemente, o que se tem são muitos atributos para a descrição da ocorrência dos casos. Como o número de atributos é diretamente proporcional à dificuldade de analisar os dados, técnicas de redução de dimensionalidade configuram ferramentas importantes para promover uma maximização dos resultados da análise. Entre os exemplos de técnicas de redução de dimensionalidade estão a "Análise de Componentes Principais" (técnica estatística) e a Fastmap [Fal 1995], que buscam identificar os atributos mais relevantes.

2.1.1.2 Caracterização dos Dados

Num processo de visualização, a determinação de qual técnica deve ser empregada em uma determinada aplicação ou situação merece bastante atenção. Uma caracterização dos dados seria a consideração inicial na escolha de uma técnica de visualização. Na tentativa de enquadramento de aplicações em técnicas, alguns autores propõem certas classificações. Shneiderman [Shn 1996], por exemplo, classificou as técnicas segundo os tipos de dados e as tarefas a serem realizadas pelo usuário. Segundo ele, os dados podem ser: temporais, unidimensionais (1D), bidimensionais (2D), tridimensionais (3D), multidimensionais (nD), dirigidas à visualização de hierarquias e de relacionamentos (grafos). Freitas e Wagner [Fre

1995] apresentam uma proposta de caracterização de dados baseada em critérios como: classe (tipo) de informação, tipos de valores, e natureza e dimensão do domínio (vide resumo na *tabela 2.2*).

***Tabela 2.2** – Sumário da caracterização de dados, exemplos de domínios diferentes [Fre 2001]*

Critério	Classe *	Exemplo
Classe de Informação	Categoria	Gênero
	Escalar	Temperatura
	Vetorial	Grandezas físicas associadas à dinâmica de fluidos
	Tensorial	
	Relacionamento	Link num hiperdocumento
Tipo de Valores	Alfanumérico	Gênero
	Numérico (inteiro, real)	Temperatura
	Simbólico	Link num hiperdocumento
Natureza do Domínio	Discreto	Marcas de automóveis
	Contínuo	Superfície de um terreno
	Contínuo-discretizado	Anos (tempo discretizado)
Dimensão do Domínio	1D	Fenômeno ocorrendo no tempo
	2D	Superfície de um terreno
	3D	Volume de dados médicos
	n-D	Dados de uma população

* Classe do Dado

De forma geral, os valores assumidos pelas variáveis podem ser classificados nos formatos básicos ***nominal*** e ***quantitativo***. O primeiro apresenta valores claramente distintos, discretos e enumeráveis. O segundo representa valores numéricos, contínuos, sobre os quais podem ser aplicadas operações aritméticas. Os dados ***nominais*** podem ser *categóricos*, em que os valores não têm uma relação de ordem (ex.: "verde", "vermelho", azul"; "GM", "Fiat", "Ford"), ou *ordinais*, apresentando relação de ordem (ex.: "segunda", "terça", "quarta"; "básico", "inter mediário", "avançado"). Por sua vez, os dados **quantitativos** podem ser *intervalos*, nos quais os valores são ordenados e medidos em unidades fixas e iguais (ex.: ano); podem ser uma *razão*, em que os valores são ordenados em um escala de medidas na qual é definido inerentemente um valor de referência zero (ex.: distâncias); e alguns *sub-tipos* particulares, tais como datas, horas e coordenadas espaciais.

Essa caracterização mais genérica, torna mais compreensível como podem ocorrer as transformações nos dados apontadas no modelo de referência de Card et al. Por exemplo, uma variável originalmente quantitativa, tal como o tempo de duração de um filme [0, 360], pode ter

seus valores separados em faixas como (“curto”, “médio”, “longo”), denotando agora um tipo de variável nominal/ordinal.

2.1.2 Técnicas para Visualização Exploratória de Dados Multidimensionais

Como visto anteriormente, ***Visualização de Informação*** pode ser entendida como uma extensão das aplicações científicas de Visualização, atendendo a uma tendência do uso de gráficos potentes para permitir a interpretação de informações complexas e a dedução de novos conhecimentos, explorando a natural capacidade de percepção do ser humano [Fre 2001]. Aplicando-se em inúmeras atividades profissionais, acadêmicas e de pesquisa, Visualização de Informação denota o conjunto de técnicas usadas para mapear graficamente informações abstratas multidimensionais - de natureza não gráfica e não necessariamente geradas por um fenômeno físico - armazenadas em grandes bases de dados [Kei 1996]. Ainda assim, encontrar formas de representação de grandes quantidades de dados multidimensionais, centradas no ser humano, isto é, capazes de efetivamente auxiliar os usuários no processo interativo de análise e interpretação [Kei 1994], permanece um desafio.

Antes mesmo do uso de computadores para criar visualizações, a visualização de dados de duas ou três dimensões já era realizada, e suas técnicas têm sido usadas por muitos anos, como bem ilustrado por Tufte [Tuf 1983; Tuf 1990]. Quando computadores começaram a ser usados para criar visualizações, também começou o desenvolvimento de muitas técnicas novas, bem como a extensão de técnicas existentes, para que trabalhem com grandes volumes de dados e permitam interação. Porém, para a maioria dos dados armazenados em base de dados, não há um mapeamento adequado no sistema de coordenadas cartesianas, visto que os dados não têm uma semântica inerente no espaço bidimensional ou tridimensional. Dessa forma, bases de dados relacionais são consideradas, genericamente, como conjuntos de dados multidimensionais, com os atributos da base de dados correspondendo às dimensões, ou variáveis [Kei 1996].

Keim e Kriegel [Kei 1996] e Wong e Bergeron [Won 1997] enumeram técnicas bem conhecidas para visualização de conjuntos de dados multidimensionais e as descrevem segundo critérios diferenciados, ainda que com alguns elementos em comum. Assim, as técnicas de

Visualização de Informação recebem diferentes taxonomias, variando entre autores. Ainda em [Kei 1996], os autores agruparam as técnicas de exploração visual de dados em seis categorias, segundo a abordagem adotada para o mapeamento dos dados em uma representação visual: Projeções Geométricas, Orientadas a *Pixels*, Iconográficas, Hierárquicas, Baseadas em Grafos e Híbridas.

Em algumas técnicas, a representação visual reflete diretamente características próprias dos dados. Exemplos disto são as técnicas hierárquicas *Cone Tree* [Rob 1991] e *TreeMap* [Joh 1991], em que a própria natureza dos dados apresenta uma correlação explícita entre níveis e/ou subconjuntos (ex.: estruturas de diretórios em sistemas de arquivos; ligação entre documentos de um *site* da Web), embora essas técnicas sejam aplicáveis também a dados que não apresentam natureza hierárquica. Por outro lado, as técnicas podem justamente apontar relações que estão implícitas nos dados. As próximas seções exemplificam técnicas representativas de cada categoria, como forma de ilustrar as principais características de cada grupo. A exceção está na categoria denominada "híbrida", que configura justamente uma mescla de características dos outros grupos de técnicas.

2.1.2.1 Coordenadas Paralelas: Projeções Geométricas

Em Projeções Geométricas, o objetivo é identificar projeções de interesse em conjuntos de dados multidimensionais. Em particular, uma técnica bastante utilizada denomina-se *Coordenadas Paralelas* (*Parallel Coordinates*). Inicialmente apresentada por Inselberg [Ins 1985] como uma técnica de Geometria Computacional, e posteriormente contextualizada em Visualização de Informação [Weg 1990], Coordenadas Paralelas destaca-se justamente pela perspectiva multidimensional conferida à representação visual. Nela, um espaço de dimensão *k* é mapeado para um espaço visual bidimensional, usando *k* eixos eqüidistantes e paralelos a um dos eixos principais (x ou y). Cada eixo representa uma dimensão do conjunto de dados, sobre o qual é mapeado linearmente, do menor ao maior, o intervalo de valores de dados correspondente. Cada item de dado é exibido como uma linha poligonal que intercepta cada eixo no ponto correspondente ao valor do atributo associado ao eixo. A *figura 2.2* apresenta um esquema básico de Coordenadas Paralelas e a *figura 2.3* exemplifica seu uso por meio da implementação em uma aplicação.

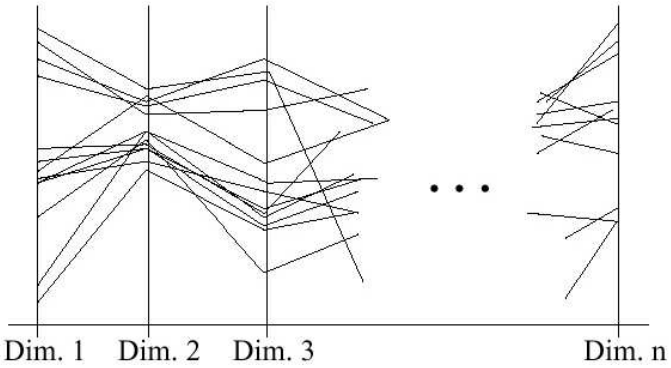

***Figura 2.2** – Coordenadas Paralelas [Kei 1996]*

Visto que gera uma representação planar, essa técnica transforma relações multivariadas em padrões bidimensionais [Weg 1996; Ins 1997], os quais permitem que sejam identificadas características como diferenças na distribuição de dados e correlações entre atributos. A *figura 2.3* ilustra a presença dessas características: no eixo rotulado "*precipitation*", vê-se uma concentração de linhas em um determinado intervalo de valores; correlações entre atributos podem ser identificadas, a exemplo das de formato X (cruzamento de linhas) entre os eixos "*sea level*" e "*humidity* 700 hPa" (eixos 3 e 4), denotando a existência de uma correlação inversa entre os atributos correspondentes [Ins 1997].

Embora simples, a técnica Coordenadas Paralelas mostra-se poderosa para identificar diferentes distribuições de dados e dependência funcional entre atributos. Em contrapartida, apresenta problemas como a sobreposição de linhas para grandes volumes de dados (*figura 2.4*) e, por conseguinte, uma baixa quantidade de itens de dados que podem ser apresentados simultaneamente sem a ocorrência de congestionamento visual (cerca de 1.000 itens) [Kei 1996].

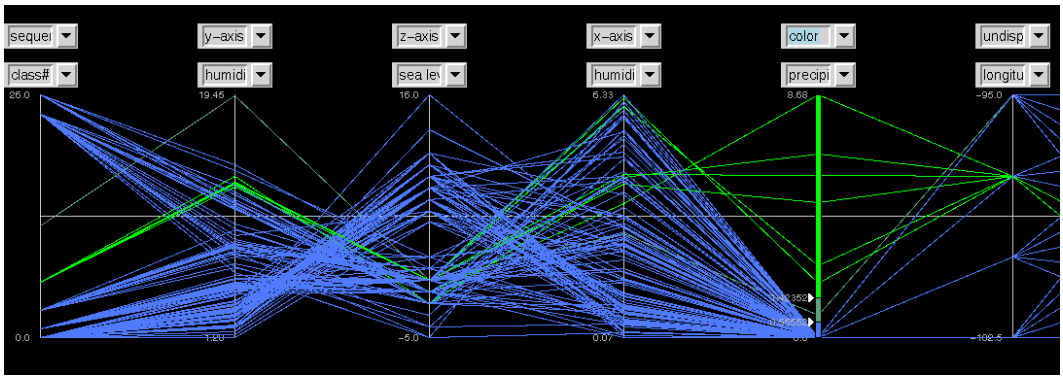

Figura 2.3 *– Exemplo de seleção de itens em aplicação com Coordenadas Paralelas [Eds 1999]*

No intuito de minimizar esse problema ou evitar a ocorrência de outros, espera-se de implementações da técnica algumas funcionalidades. Um exemplo simples é o uso de cores para destacar/selecionar itens de dados, como se pode observar na *figura 2.3*. Outras funcionalidades desejáveis, incluindo recursos de interação adequados, seriam [Ale 1998]: visibilidade (esconder ou não certos itens ou dimensões); permutação entre dimensões (no caso, eixos); dimensões re-escaláveis; zoom (para promover melhor visualização de faixas de conjuntos de dados de maior interesse); informação de dados sob demanda; múltiplas visualizações simultâneas, inclusive com acoplamento de outras técnicas; animações; aplicação de diferentes tipos de operações interativamente e/ou automaticamente (adotando-se, por exemplo, *scripts*[6]); etc.

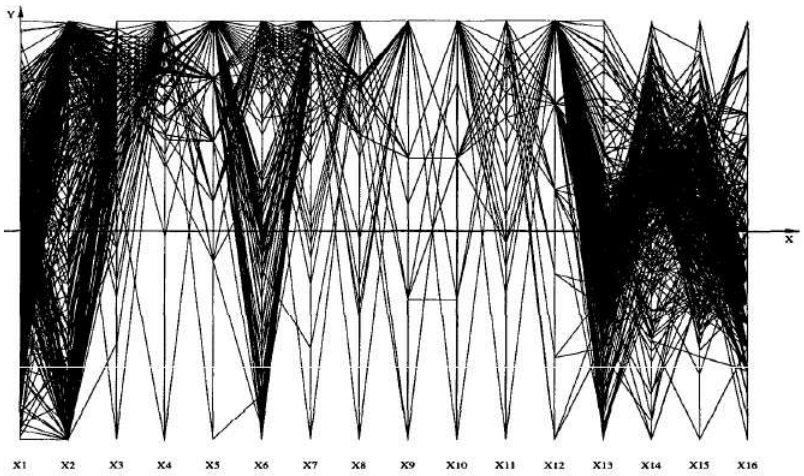

Figura 2.4 *– Sobreposição de linhas com apresentação de 473 itens [Ins 1997]*

[6] Entende-se por *script*, uma seqüência de operações previamente definida, estabelecendo um roteiro de execuções.

2.1.2.2 Técnicas Orientadas a Pixels

Em técnicas baseadas em *pixels*, a idéia básica consiste em usar um *pixel* para representar cada valor de atributo, colorindo-o conforme um mapa de cores previamente fixado de acordo com a faixa de possíveis valores do atributo, sendo que cada um desses atributos tem sua representação visual exibida em sub-janelas individuais na visualização [Kei 1994a; Kei 1996]. Para conjuntos de dados com *m* atributos, a tela é dividida em *m* janelas, como ilustrado na *figura 2.5*. Correlação e dependência funcional podem ser detectadas pela análise de regiões correspondentes nas múltiplas janelas [Kei 1996; Kei 2000], a exemplo da correlação observada entre as dimensões *MinAngle* e *RightAngle* na *figura 2.6*.

Se um único atributo for apresentado em uma janela com resolução de 1280x1024, é possível exibir mais de um milhão de valores simultaneamente. E essa é uma das vantagens desse tipo de técnica: a grande quantidade de informação que pode ser exibida simultaneamente.

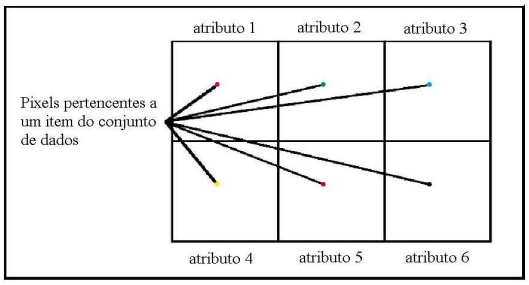

***Figura 2.5** – Arranjo de janelas de atributos para casos com seis dimensões [Kei 2000]*

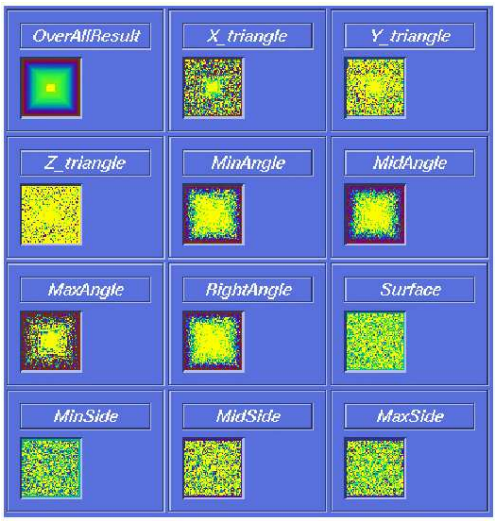

***Figura 2.6** – Identificação de correlação e dependências funcionais no VisDB [Kei 1996]*

Para o sucesso na aplicação desse tipo de técnica, alguns aspectos precisam receber atenção [Kei 1994; Kei 1994a; Kei 1996; Kei 2000]:

- *Arranjo dos pixels nas janelas*. Visto que cada valor de atributo é representado por um *pixel*, uma questão fundamental é como "arranjar", ou seja, a forma de distribuir os *pixels* na tela. De fato, "arranjar" é mapear um conjunto de dados unidimensionais (um atributo) nas duas dimensões da tela. Tal mapeamento deve ser executado de maneira a fornecer boas propriedades de agrupamento (*clustering*), e mantendo algum significado semântico, favorecendo a percepção de relações existentes entre os dados. Dessa forma, as técnicas recebem diferentes arranjos segundo diferentes propósitos e condições.

Quando se quer visualizar um grande conjunto de dados, pode-se utilizar as técnicas ditas "*query-independent*", que ordenam todo o conjunto de dados de acordo com algum(ns) atributo(s) e adotam um preenchimento de tela padrão (esquerda-direita/cima-baixo) para o arranjo dos valores de dados. Tais técnicas são úteis especialmente para dados com uma ordenação natural inerente, como séries temporais, podendo ser visualizados diretamente de uma base de dados. A técnica denominada "Padrões Recursivos" (*figura 2.7*) baseia-se num esquema recursivo genérico, que permite que padrões de baixo nível sejam usados como blocos para construir padrões de mais alto nível, constituindo uma série de níveis de padrões. Como exemplo, numa série de dados temporais, em que se executaria diversas coletas diárias de um mesmo parâmetro, primeiramente seria desejável que os -dados de um dia sejam agrupados (arranjados proximamente), depois esses dias formariam semanas, depois meses, anos, e assim sucessivamente.

Quando não há a referida ordenação e o objetivo é uma exploração interativa de um banco de dados, a visualização poderá ser feita considerando a relação entre os itens de dados e o resultado de uma consulta, utilizando-se as técnicas "*query-dependent*". Nesse caso, além dos itens de dados que satisfazem a consulta, é de interesse mostrar também aqueles que se aproximam da resposta. Ao invés do *valor* do atributo, o que é exibido é uma *distância* calculada entre os dados e o(s) valor(es) estabelecido(s) na consulta, ordenados segundo uma distância global — calculada pela combinação das distâncias de cada atributo, ponderadas por um peso correspondente, que determina a relevância dada ao atributo —, de modo que os atributos referentes a um item de dado em particular ocupam a mesma posição nas suas respectivas janelas. A disposição dessa ordenação, ou seja, o arranjo dos itens de dados, centraliza os itens mais relevantes no meio da janela (itens que atendem integralmente a consulta, têm distância

global igual a zero), e arranja os itens de dados menos relevantes em direção à parte externa da janela, seguindo esquemas como espirais simples (*figura 2.8a*) ou curvas de Peano-Hilbert (*figura 2.8b*) ou Morton (*figura 2.8c*).

Técnicas que utilizam eixos com valores positivos e negativos (*figura 2.9*), nas quais cada eixo representa um atributo, acrescentam uma informação de direção das distâncias (se o valor do atributo varia para mais ou para menos). Contudo, a quantidade de itens exibidos é menor, e algumas regiões do gráfico podem ficar sem preenchimento em virtude da probabilidade da distribuição dos valores de distância, entre positivo e negativo, não ser uniforme.

- *Mapeamento de cores*. O uso de cor permite um número maior de JNDs (*Just Noticiable Differences*), se comparado à escala de cinzas. As JNDs são as cores percebidas como diferentes e configura-se em grande desafio determinar uma escala de cores que maximize o seu número, e que também seja natural para o usuário. A meta é evitar que relações entre atributos fiquem ocultas e que artefatos visuais, que induzem à interpretações equivocadas, sejam introduzidos.

- *Formato das janelas*. Consoante com o formato da tela do computador, o formato retangular para as janelas garante um bom uso de espaço desprezível. Todavia, o estabelecimento de relações entre atributos pode ser dificultado pela distância relativa entre as janelas na tela. A forma circular é um formato alternativo ao retangular, adotado na técnica denominada *circle segments* (*figura 2.10a*). Cada atributo é, então, visualizado em um segmento do círculo, com valores posicionados a partir do seu centro e seguindo o caminho dado pelas *draw lines*, ortogonais às *halving lines* (*figura 2.10b*), que são as linhas que separam os segmentos.

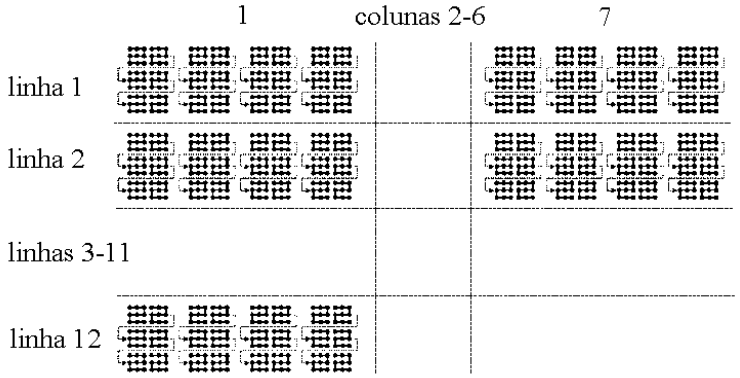

***Figura 2.7** – Esquema básicos do arranjo "Padrões Recursivos" [Kei 2000]*

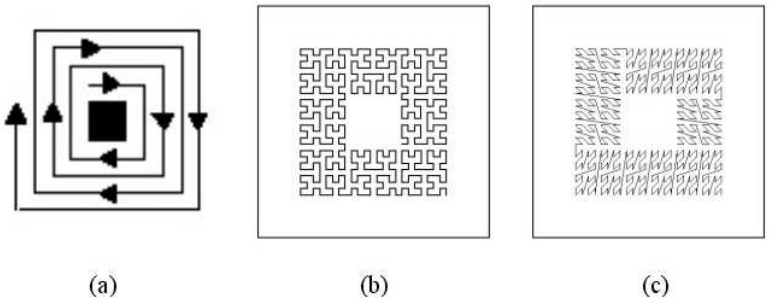

Figura 2.8 *– Esquemas de arranjo de pixels: (a) em espiral simples [Kei 1994], com curvas de (b) Peano-Hilbert e (c) Morton [Kei 1996]*

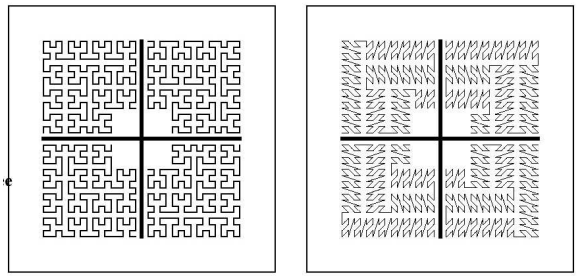

Figura 2.9 *– Esquema com eixos para identificar distâncias positivas e negativas [Kei 1996]*

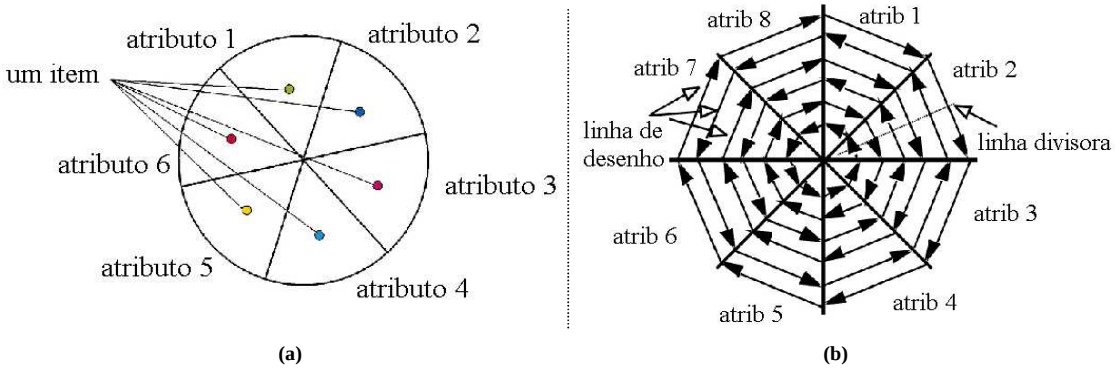

Figura 2.10 *– (a) Formato circular para distribuição dos pixels; (b) Ordenação dos pixels em cada segmento [Kei 2000]*

2.1.2.3 Stick Figures: Iconográficas

As técnicas dessa classe têm como principal característica o uso de ícones como forma de mapear os valores dos atributos de um item de dado multidimensional. Cada característica visual do ícone corresponde a um atributo.

Apresentada por Pickett e Grinstein [Pic 1988], a *Stick Figures* utiliza as duas dimensões da tela para mapear duas dimensões dos dados e as demais dimensões são mapeadas para os ângulos e/ou comprimentos de segmentos de um ícone formado por múltiplos segmentos de reta. A *figura 2.11a* apresenta um ícone com uma configuração que apresenta 5 variáveis, na qual uma dimensão é mapeada pela inclinação do corpo do ícone, e a orientações das varetas permitem mapear outras quatro dimensões. Uma família de *Stick Figures* é ilustrada na *figura 2.11b*, em que cada uma tem um corpo e 4 segmentos. Outras formas de representar dimensões nesses ícones seriam por meio da variação de cores e espessuras das varetas.

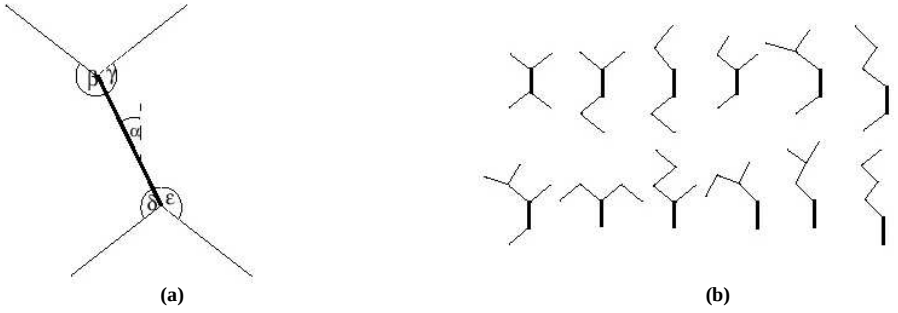

***Figura 2.11** – (a) Ícone que representa 5 variáveis; (b) Família de ícones [Won 1997]*

Quando mapeados na tela, os ícones (um para cada item de dado) formam texturas que variam de acordo com as características dos dados, permitindo a exploração da capacidade humana de perceber e distinguir texturas em imagens complexas [Won 1997], ao passo que os padrões percebidos na imagem podem, então, indicar dependência funcional entre os atributos visualizados [Kei 1996; Won 1997]. A *figura 2.12* exibe uma imagem formada pela plotar ícones que representam 5 variáveis, exemplificando como essas texturas podem ser formadas.

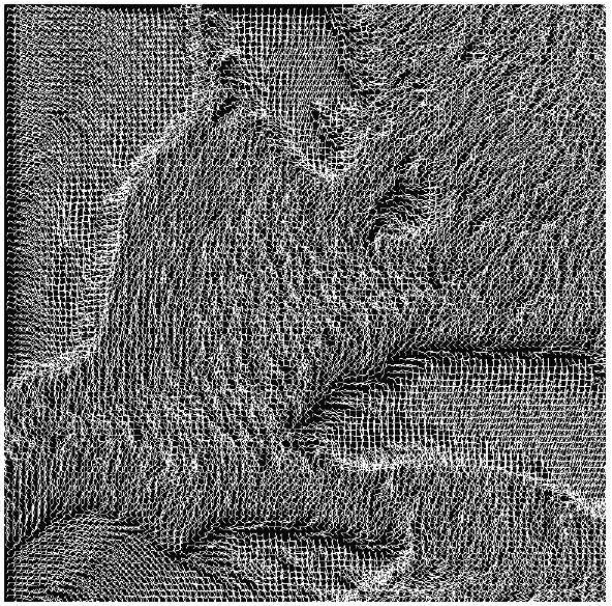

***Figura 2.12** – Imagem da Região dos Grandes Lagos com ícones de 5 dimensões [Ank 2001]*

Dependendo da configuração adotada, essa técnica consegue demonstrar grande quantidade de itens de dados, mas há limitações quanto à quantidade de dimensões que podem ser mapeadas sem que haja detrimento na capacidade de representar características detectáveis dos dados. Outra problemática diz respeito ao fato de que o reconhecimento de um importante padrão na imagem depende da seleção de um mapeamento adequado dos parâmetros dos dados para os parâmetros visuais. O número desses possíveis mapeamentos visuais cresce em ordem fatorial em relação ao número de dimensões mapeadas, podendo tornar-se um grande gargalo no processo de visualização [Won 1997].

2.1.2.4 Dimensional Stacking: Hierárquicas

Nas técnicas hierárquicas de visualização, o espaço n-dimensional dos dados (não necessariamente de natureza hierárquica) é dividido em subespaços que são organizados e exibidos na forma de hierarquia, projetando ou embutindo esses espaços uns dentro dos outros.

Na técnica *Dimensional Stacking*, o espaço n-dimensional (discreto) é subdividido em espaços bidimensionais. Uma das maiores vantagens da *Dimensional Stacking*, em relação a

outras técnicas hierárquicas, é que ela não precisa de funções ou regras extras para que se possa plotar os dados na representação [Kei 1996] [Won 1997].

A *figura 2.13* um esquema conceitual da Dimensional Stacking com um mapeamento de quatro atributos. A *figura 2.14* apresenta um exemplo de sua aplicação no conjunto de dados "*Iris Plant Flower*", em que: cada cor representa um tipo de flor (com alguns quadrados com classificações mistas); no nível mais externo estão representados os comprimentos das pétalas (eixo-x) e sépalas (eixo-y), determinantes para a classificação da flor; e, no nível mais interno, estão as medidas de altura das mesmas partes da flor, seguindo a mesma orientação dos eixos.

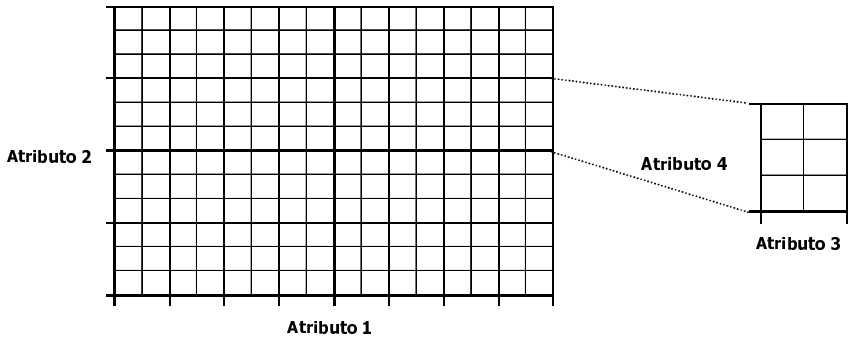

Figura 2.13 *– Modelo Conceitual da Técnica Dimensional Stacking [Ank 2001]*

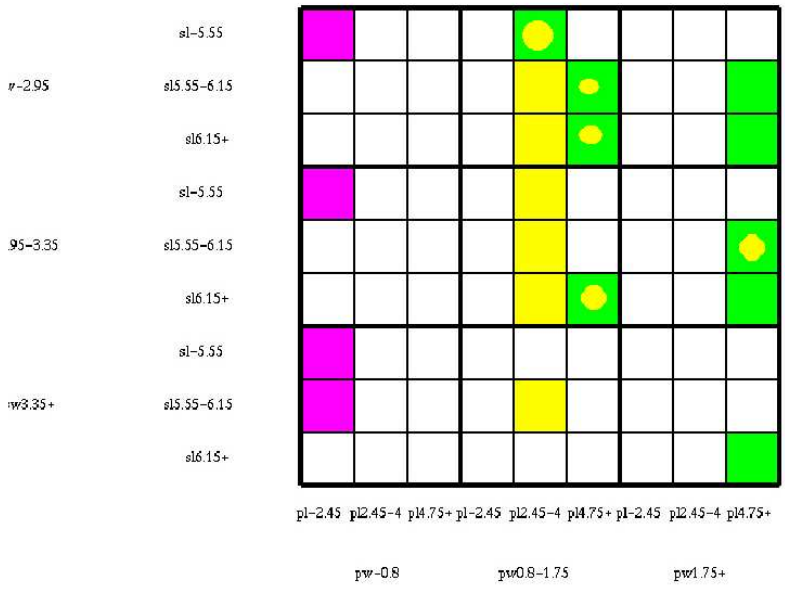

Figura 2.14 *– Visualização da Dimensional Stacking aplicada ao conjunto de dados "Iris Flower" [Hof 1999]*

O uso dessa técnica é particularmente interessante na detecção de agrupamentos, de pontos com comportamentos discrepantes, e de padrões [Hof 1999]. Todavia, há limitações referentes ao número de atributos a serem visualizadas (acima de 9 torna-se bastante difícil) e ao número de valores distintos que um mesmo atributo pode assumir. Além disso, a eficácia do processo de análise depende diretamente do arranjo hierárquico dos atributos (normalmente, os atributos mais importantes devem estar posicionados mais externamente) e dos critérios de categorização dos dados [Kei 1996; Won 1997; War 1994]. Essa última restrição é uma limitação artificial imposta pelo dispositivo gráfico, que pode ser minimizada pelo uso de técnicas de interação adequadas.

2.1.2.5 Técnicas Baseadas em Grafos

A idéia básica dessa categoria é visualizar grafos volumosos usando técnicas que mapeiem as características (direcionado/não-direcionado, cíclico/acíclico, etc.) de um dado grafo, de maneira clara e rápida. As técnicas podem ser subdivididas segundo a dimensionalidade visual da representação: 2-D ou 3-D.

As visualizações geradas dependem de muitos fatores que refletem características inerentes aos grafos, muitas delas refletindo definições próprias da teoria dos grafos. Em relação ao desenho de grafos 2-D tem-se como propriedades (existentes ou não): a planaridade (não cruzamento de linhas – arestas), a ortogonalidade (somente linhas ortogonais) e a propriedade de distribuição em grade (coordenadas dos vértices são números inteiros). As propriedades estéticas, objetivando a otimização são: número mínimo de cruzamentos, exibição ótima de simetria, exibição ótima de agrupamentos, número mínimo de curvas em grafos com poli-linhas, distribuição uniforme dos vértices, e comprimento uniforme das arestas. A *figura 2.15* exemplifica o uso das técnicas para representação de grafos em 2-D, adequando-se as propriedades descritas acima ao objetivo (tipo) do grafo. A *figura 2.16* apresenta uma representação em 3-D.

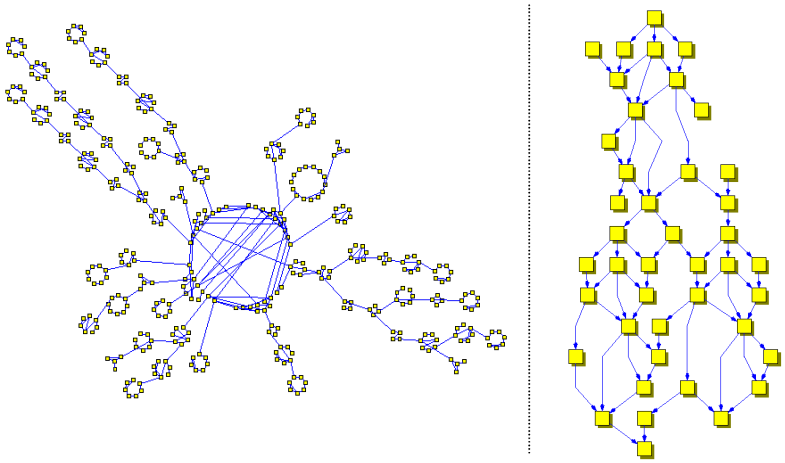

Figura 2.15 *– Exemplos de aplicações de Visualizações 2D de grafos: (a) Grafo otimizado para agrupamentos; (b) Grafo acíclico não-direcionado [Ank 2001]*

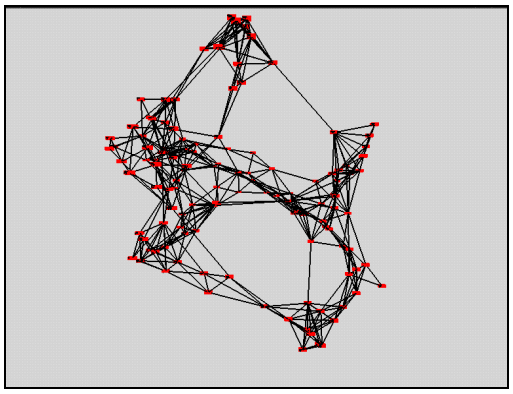

Figura 2.16 *– Exemplo de aplicação de Visualização 3D de grafos: Grafo Otimizado para agrupamentos [Ank 2001]*

2.2 Extração de Conhecimento e Mineração de Dados

Segundo Fayyad et al. [Fay 1996; Fay 1998], KDD é o processo global de identificar nos dados um padrão, um modelo ou uma estrutura válida, nova, potencialmente útil e interpretável. Embora os termos *Descoberta de Conhecimento em Bases de Dados* (*Knowledge Discovery in Databases* — KDD) e *mineração de dados* (*data mining* — DM) sejam muitas vezes considerados como tendo o mesmo significado, aqueles autores estabelecem uma clara distinção

entre esses dois conceitos. DM é apontado como um passo particular na seqüência de passos envolvidos no processo de descoberta de conhecimento (*figura 2.17*). Como estabelecido no primeiro *workshop* da área, em 1989 ([Pia 1991] apud [Fay 1998]), o termo KDD enfatiza que o conhecimento (*knowledge*) é o produto final de um processo de descoberta guiado por dados.

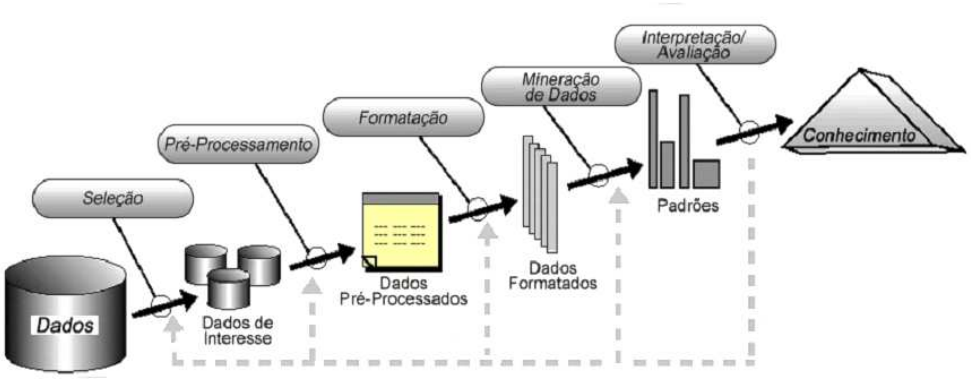

Figura 2.17 *– Passos do Processo de KDD [Fay 1996]*

Como principal motivo para a "confusão" entre os termos KDD e DM, tem-se que a mineração de dados é encarada como o núcleo do processo de KDD, sendo responsável pelo mapeamento dos dados para algum tipo de informação ou modelo. Salienta-se, porém, que mesmo sendo dada grande atenção às técnicas de DM, que podem representar entre 15% e 25% do esforço do processo de descoberta [And 1999], figura-se a importância de todos os passos para o sucesso do processo como um todo [Fay 1996].

A validade e o grau de interesse das muitas informações que podem ser extraídas dos conjuntos de dados são dependentes do domínio de aplicação e do usuário [Fay 1998]. O entendimento desse domínio, o levantamento de conhecimento prévio e a definição precisa do objetivo do usuário são aspectos importantes para a condução bem sucedida do processo de descoberta, influenciando, por exemplo, a escolha do método de DM a ser utilizado.

Além da relevância desses aspectos, o processo de descoberta de conhecimento apresenta como características importantes a interatividade e a iteratividade. Essas não requerem, necessariamente, uma seqüência pré-definida, e compõem o processo por repetições e experimentações ao longo de diversas etapas, exigindo a tomada de muitas decisões pelo usuário. A *figura 2.17* esquematiza os passos componentes do processo de KDD. Esse percurso que

transforma dados brutos em informação útil, ou conhecimento, caracteriza-se pelas diversas formas de processamento aplicadas aos dados (passos do KDD) [Fay 1996]: seleção, pré-processamento, transformação, mineração de dados e Interpretação/Avaliação.

Segundo Keim e Kriegel [Kei 1996], o passo denominado mineração de dados envolve a identificação de subconjuntos de um conjunto de dados e de hipóteses sobre os mesmos. É importante salientar que, para que as ferramentas de DM auxiliem a identificação desses dois elementos, pode não ser tão importante especificar formalmente as hipóteses e os contextos. DM pode ser interpretado, assim, como um *processo interativo de geração de hipóteses*. De forma simplificada, DM, é o processo de pesquisar e analisar grandes volumes de dados, identificando estruturas regulares e irregulares [Kei 1994], constituindo mais um exercício *indutivo* do que hipotético-dedutivo [Han 1998]. Em função do tipo de informação extraída dos dados, técnicas de DM podem ser agrupadas em [Che 1996]: Regras de Associação, Generalização e Sumarização de Dados, Classificação de Dados, Agrupamento (*Clustering*) de Dados e Pesquisa por Similaridade Baseada em Padrões.

Resta salientar que DM e KDD são campos interdisciplinares, nos quais avanços significativos requerem o uso de técnicas de diversas áreas [Fay 1998], tais como Visualização, Inteligência Artificial, Banco de Dados, Estatística e Computação de Alto Desempenho, as quais compartilham o objetivo de extração de conhecimento de alto nível a partir de conjuntos volumosos de dados brutos. As abordagens de KDD focalizam na extração de modelos dos dados sem necessariamente uma hipótese formulada *a priori*, diferentemente de abordagens tradicionais em Estatística. Focalizam, também, conjuntos com maior volume de dados do que os tradicionalmente manipulados pelas técnicas estatísticas. Em conseqüência, uma propriedade importante nos algoritmos é a 'escalabilidade' [Fay 1996; Fay 1998], ou seja, a capacidade de continuar operando correta e adequadamente, à medida que o tamanho dos arquivos de dados aumenta. Questões pertinentes à localização, relação e organização dos dados, e da presença de dados com valores inesperados ou ausência de informação, podem inviabilizar a aplicação direta de métodos estatísticos clássicos [Han 1998], acabando por motivar a criação de novos métodos, bem como o desenvolvimento de técnicas integradas para um DM eficiente e eficaz [Che 1996].

2.3 Integração de Visualização e Mineração de Dados

Atualmente, as técnicas para visualização de dados são tidas como instrumentos indispensáveis ao processo de DM [Rez 2003]. Visualização tem apresentado função relevante nas tarefas de DM, como Visualização de Modelos de DM e Exploração Visual dos Dados. Enquanto naquela estão as técnicas visuais para produzir descoberta de conhecimento entendível e interpretável por humanos, nessa estão caracterizadas a exploração interativa de conjunto de dados utilizando-se representações gráficas, sem uma forte dependência de hipóteses e modelos, na tentativa de identificar padrões de interesse não conhecidos previamente.

Keim [Kei 2001] define que uma exploração visual dos dados pode ser vista como um processo de geração de hipóteses, segundo o qual a visualização dos dados permite ao usuário adquirir percepções dos dados, podendo provocar o surgimento de novas hipóteses, que, por sua vez, podem também ser confirmadas ou rejeitadas com o uso da exploração visual. Além disso, ele acrescenta que, comparada a técnicas automáticas de mineração de dados em estatística e máquinas de aprendizado, a exploração visual dos dados apresenta vantagens excedentes: lida mais facilmente com dados altamente heterogêneos e ruidosos; é intuitiva; e não requerer maior entendimento de complexos algoritmos ou parâmetros matemáticos ou estatísticos.

A integração de técnicas de Mineração de Dados e Visualização é referenciada na literatura como *Mineração Visual de Dados* (*Visual Data Mining* — VDM) [Gan 1996; Kei 1996; Won 1999]. São utilizados também outros termos, como *Discovery Visualization* (DV) [Rib 1999] e Análise Visual [Roh 1999]. Em VDM, o processo de análise é reforçado pelas vantagens oriundas da interação direta com o usuário e da orientação do processo pelo usuário, obtidas com o uso de técnicas de visualização [Kei 2000]. Sobretudo, quando técnicas de DM requerem grande interação com o usuário, e essa interação se mostra bastante complexa, técnicas de Visualização podem certamente ser exploradas para dar suporte ao processo de decisão, deixando de caracterizarem-se apenas como técnicas separadas de DM e Visualização, para configurarem-se em técnicas de VDM [Won 1999].

Segundo Wong, existem, basicamente, duas formas de integrar Visualização e DM [Won 1999]: Acoplamento Forte, em que a Visualização e o processo analítico são integrados em uma única ferramenta, aproveitando os pontos fortes de cada uma das áreas; e Acoplamento Fraco,

em que técnicas das duas áreas são simplesmente intercaladas, possibilitando um aproveitamento apenas parcial do potencial de cada uma. Com o uso de acoplamento forte, poder-se-ia ter, por exemplo, a tomada de decisões humanas guiadas por representações visuais, no lugar de alguns passos matemáticos executados de forma automática pelo algoritmo que implementa o processo analítico. Algumas avaliações experimentais de acoplamento forte que combinam recursos de Visualização a algoritmos de agrupamento [Hin 1999], mostram que a combinação de técnicas visuais e automáticas melhora consideravelmente a eficiência do processo de DM e estimula um melhor entendimento dos resultados.

Como visto na *seção 2.2*, o processo formado pelos passos do KDD apresenta como característica uma grande **iteratividade**. A **interatividade**, estaria normalmente restrita aos passos iniciais do processo. Na visão de Ankerst e Keim [Ank 2001], o processo clássico de KDD pode ser estendido para que o usuário possa introduzir seu conhecimento do domínio em todos os passos do processo de KDD, no que foi chamado de "Processo de Descoberta de Conhecimento Centrado no Usuário" (*figura 2.18*).

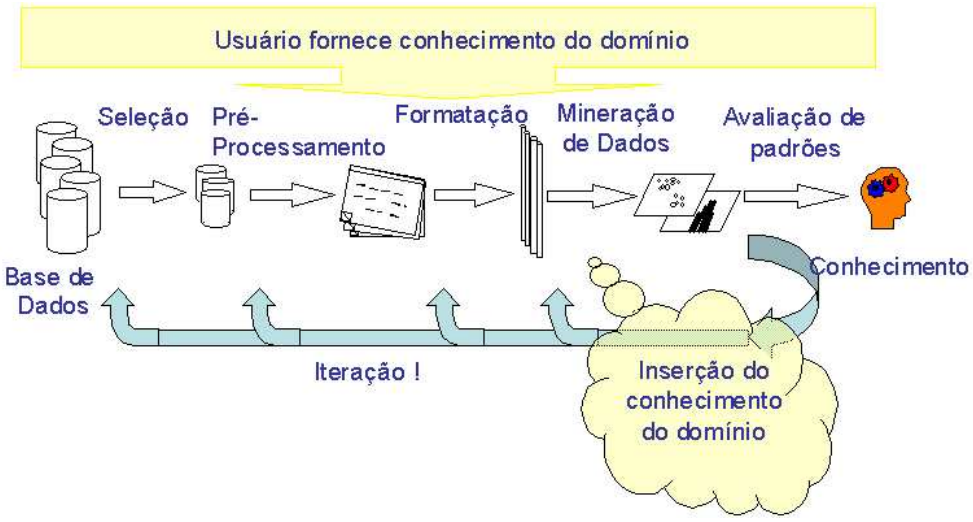

Figura 2.18 *– Processo de KDD Centrado no Usuário [Ank 2001]*

Ankerst [Ank 2000] definiu Mineração Visual de Dados como sendo "um passo no processo de KDD que utiliza Visualização como um canal de comunicação entre o computador e o usuário, para produzir padrões novos e interpretáveis". Nessa abordagem, a Visualização seria principalmente empregada nas duas fases essenciais do processo: mineração de dados e

avaliação. Mineração de Dados, então, passa a ser um dos passos em que o usuário pode introduzir seu conhecimento de domínio no processo de KDD, ao invés de ser um passo meramente automatizado.

No intuito de promover uma melhor definição de VDM, Ankerst propôs também uma classificação de como as diversas abordagens de Visualização podem ser integradas ao processo de KDD. Essa classificação baseia-se na seqüência (em que momento) e no equilíbrio entre as partes automáticas e interativas (visuais) do processo. São elas:

- *Visualização dos dados*: os dados são visualizados sem prévia execução de algoritmos sofisticados. Por meio de interação e operações na visualização, o usuário tem total controle sobre a busca no espaço de busca, focalizando-o e/ou delimitando-o.

- *Visualização de resultados intermediários de uma mineração*: ocorre quando os algoritmos que executam uma análise dos dados não produzem padrões finais, mas sim, padrões intermediários que podem ser visualizados. Dessa maneira, o usuário pode encontrar padrões de interesse na visualização, visto que essa promove um conhecimento do domínio e, por conseguinte, promove também um direcionamento da busca. Essa abordagem torna-se particularmente importante quando se observa que não há algoritmos genéricos para mineração de dados, e que pode fornecer uma forma eficiente de avaliar e validar o andamento do processo.

- *Visualização de resultados da mineração*: corresponde às visualizações subseqüentes à extração de padrões nos dados, tornando-os mais facilmente interpretáveis. Além disso, baseado nessa visualização, o usuário pode querer retornar ao algoritmo de mineração de dados e reexecutá-lo utilizando diferentes parâmetros de entrada.

A *figura 2.19* ilustra como se dá a integração de técnicas de Visualização no processo de mineração, segundo a proposta de Ankerst e Keim [Ank 2001]. Nela, observa-se que a "Visualização Fortemente Integrada" determina uma a bordagem em que a(s) técnica(s) de visualização permeiam todo o processo de extração de padrões.

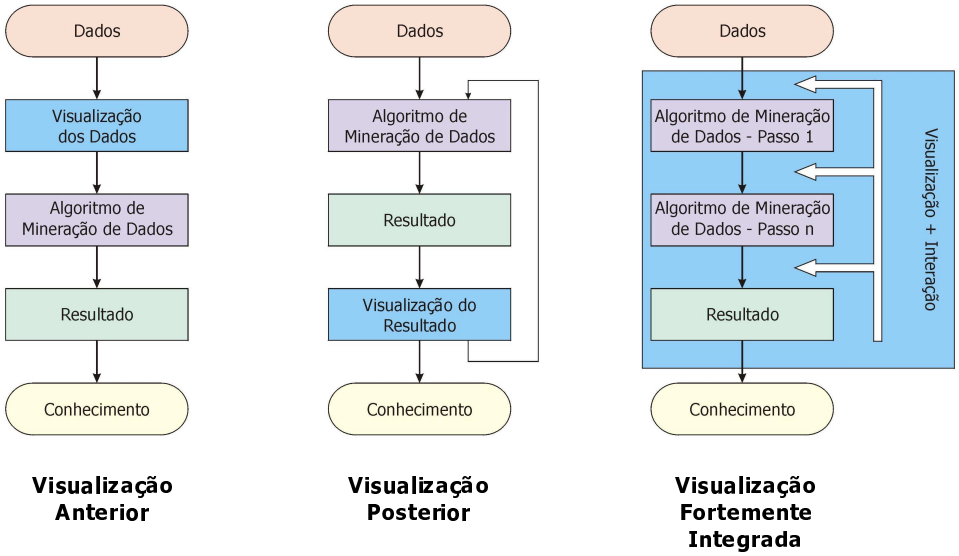

Figura 2.19 *– Categorias de Mineração Visual de Dados [Ank 2001]*

2.4 Sumário

Esse capítulo tratou de aspectos envolvidos com Visualização e Mineração de Dados, e da atual tendência de tentar integrá-los, originando a chamada Mineração Visual de Dados.

A Visualização pode ser descrita por meio de modelos de referência que descrevem o processo de mapear dados e informações em um formato gráfico, integrando o usuário ao longo de todo esse processo. Quando as informações a serem mapeadas são abstratas, a Visualização é chamada de Visualização de Informação. Os conjuntos de dados apresentam características essenciais para a escolha das técnicas de visualização. Essas, por sua vez, podem ser agrupadas em categorias que definem os principais aspectos inerentes às técnicas aí inseridas.

Mineração de Dados é um passo particular da Descoberta de Conhecimento em Bases de Dados, o qual corresponde a um processo que transforma dados brutos em informação útil. Mineração de dados é considerada como o núcleo desse processo, representando a pesquisa e análise de grandes volumes de dados, identificando estruturas regulares e irregulares.

Mineração Visual de Dados mostra-se particularmente útil quando o processo de mineração requer grande integração com o usuário e essa integração se mostra bastante

complexa. Como Visualização pode ser explorada para integrar o usuário por meio de representações visuais, mineração de dados pode ser claramente beneficiada por um ambiente de descoberta que integre as capacidades únicas do ser humano de percepção ao poder de processamento dos computadores.