JULIANA KEIKO YAMAGUCHI

DIRETRIZES PARA A ESCOLHA DE TÉCNICAS DE VISUALIZAÇÃO APLICADAS NO PROCESSO DE EXTRAÇÃO DO CONHECIMENTO

MARINGÁ

2010

JULIANA KEIKO YAMAGUCHI

DIRETRIZES PARA A ESCOLHA DE TÉCNICAS DE VISUALIZAÇÃO APLICADAS NO PROCESSO DE EXTRAÇÃO DO CONHECIMENTO

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual de Maringá, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientadora: Profª. Drª. Maria Madalena Dias

MARINGÁ

2010

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR., Brasil)

Yamaguchi, Juliana Keiko
Y19d Diretrizes para a escolha de técnicas de visualização aplicadas no processo de extração do conhecimento. / Juliana Keiko Yamaguchi. -- Maringá, 2010.
160 f. : il. color., figs., tabs.

Orientador : Prof.ª Dr.ª Maria Madalena Dias.
Dissertação (mestrado) - Universidade Estadual de Maringá, Programa de Pós-Graduação em Ciência da Computação, 2010.

1. Visualização de dados - Técnicas. 2. Visualização de dados - Critérios de escolha. 3. Visualização de dados - Extração de conhecimento. I. Dias, Maria Madalena, orient. II. Universidade Estadual de Maringá. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 22.ed.006.42

JULIANA KEIKO YAMAGUCHI

DIRETRIZES PARA A ESCOLHA DE TÉCNICAS DE VISUALIZAÇÃO APLICADAS NO PROCESSO DE EXTRAÇÃO DO CONHECIMENTO

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual de Maringá, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Aprovada em 08/06/2010.

BANCA EXAMINADORA

Profa. Dra. Maria Madalena Dias
Universidade Estadual de Maringá – DIN/UEM

Profa. Dra. Clélia Franco
Universidade Estadual de Maringá – DIN/UEM

Prof. Dr. Milton Hirokazu Shimabukuro
Universidade Estadual Paulista "Julio de Mesquita Filho" – DMEC/UNESP

[...] só se vê bem com o coração.
O essencial é invisível aos olhos.

Saint-Exupéry

Agradecimentos

Palavras são ineficientes para expressar sentimentos como a gratidão e o amor. Ainda assim, quero expressar minha gratidão à Deus, por me sustentar nos momentos difíceis e ter permitido que eu cumprisse mais esta etapa da vida. Agradeço, também, por Ele ter me colocado no meio de pessoas queridas, que sempre estiveram ao meu lado, às quais sou profundamente grata.

Dedico este trabalho à memória de meus avós, Kimie e Seibu, que batalharam durante toda a vida e contribuíram para a minha formação.

Agradeço aos meus pais, Tereza e Tadaaki, pelo amor dedicado a mim e ao meu irmão Bruno; aos meus tios: Jorge, Harumi, Olívia, Elena, Kyoko e Maria pelo apoio e palavras de incentivo.

Ao André, que é muito especial para mim, por me escutar, me consolar, me apoiar, me amar sempre!

À Tatiana, que desde os primeiros tempos da graduação mostra-se ainda uma grande amiga.

À grande amiga de codinome Negavan, pelo carinho, pelas palavras de incentivo e momentos de reflexão sobre o quê tem valor na vida.

Aos amigos Lúcio e Antônio, pelo aprendizado que me proporciaram e ainda o fazem até hoje.

Aos amigos que sempre estiveram na torcida por mim: Cristina, Lucas Buchala, Danieli Hautequest e Márcia Tanimoto.

Aos amigos Marco Aurélio, Maíra, André Noel, Raquel e Silvia, pela amizade e ensinamentos que valem para toda a vida.

À minha orientadora Madalena, pela dedicação e condução do desenvolvimento deste trabalho.

À professora Clélia, por compartilhar seus conhecimentos, ajudando nos momentos de incerteza que surgiram durante a pesquisa.

Ao professor Airton, pelo incentivo e apoio.

À Inês, por sua competência frente à secretaria do mestrado.

E a todos os professores do Departamento de Informática da UEM, que fizeram parte da minha formação acadêmica e também aos colegas do mestrado, que compartilharam das mesmas angústias!

Fica registrado aqui, meu sincero sentimento de gratidão a todas essas pessoas e, também, àquelas que de alguma forma deram sua contribuição nesta etapa!

Gostaria de agradecer também à Fundação Araucária, pelo apoio financeiro e à equipe de desenvolvimento do software para a construção de diagramas, Gliffy <www.gliffy.com>, por meio do qual construí alguns gráficos presentes neste trabalho.

Juliana

Resumo

A busca de informações novas e úteis em dados, que tendem a aumentar devido ao avanço tecnológico, exige métodos mais sofisticados como os desenvolvidos para a descoberta de conhecimento em banco de dados. Tais métodos podem utilizar técnicas de mineração de dados bem como técnicas de visualização, com o intuito de extrair informações e possibilitar que o usuário, conhecedor do domínio, tenha melhores condições de compreender as informações extraídas para a tomada de decisão. Desse modo, técnicas de visualização e de mineração de dados podem ser aplicadas em conjunto. Contudo, esta integração não é obrigatória, sendo possível extrair informações dos dados apenas por meio de técnicas de visualização utilizadas como ferramenta de exploração. No entanto, a definição de qual, ou quais, dentre as técnicas de visualização existentes, poderá cumprir melhor este papel durante a extração do conhecimento e, ainda, fazê-lo da melhor forma possível, exige o entendimento sobre os fatores que podem influenciar essa escolha considerando também o conhecimento sobre o domínio da aplicação. Nesse contexto, foram realizados estudos e análises de técnicas de visualização com o objetivo de estabelecer diretrizes para a escolha dessas técnicas, constituindo-se na principal contribuição deste trabalho. Para isto, foi utilizada a metodologia Teoria Fundamentada em Dados (do inglês, *Grounded Theory*), por meio da qual a definição dessas diretrizes baseou-se em parâmetros identificados durante a pesquisa, nomeados como: tipo de dado, tipo de tarefa, volume, dimensionalidade e posição dos atributos. Estes parâmetros foram associados às técnicas de visualização existentes e foram analisados tendo como base a literatura a respeito do uso dessas técnicas. Adicionalmente, estes parâmetros foram analisados sob o ponto de vista prático, aplicando-se ferramentas computacionais de visualização em bases de dados reais e fictícias.

Palavras-chave: visualização de dados; descoberta de conhecimento; técnicas de visualização; Teoria Fundamentada em Dados.

Abstract

The search for new and useful information on data, which tends to increases due to technological advancement, requires more sophisticated methods as those developed for the knowledge discovery in database. Such methods may employ data mining and visualization techniques in order to extract information and enable the domain expert to have a better understanding about the extracted information for decision making. Thus, data mining and visualization techniques can be applied together. However, this integration is not mandatory. It is possible to extract information from data just using visualization techniques as an exploitation tool. Therefore, the definition of visualization techniques, which may better fulfill this role during the extraction of knowledge and also make it the best way, requires an understanding about what can influence this choice, including the knowledge about the domain. In this context, the main contribution of this work is an analysis of visualization techniques in order to establish guidelines for the best choice of these techniques. These guidelines were based on parameters named as: data type, task type, volume, dimensionality and position of the attributes on the display. The procedure adopted to identify these parameters was based in the Grounded Theory methodology, where parameters were described and analyzed within visualization techniques, and also using computational tools in the practical application of visualization techniques on real and fictitious databases.

Keywords: data visualization; knowledge discovery; visualization techniques; Grounded Theory.

Lista de Figuras

Lista de Tabelas

Sumário

1. Introdução

As instituições vêm acumulando grande quantidade de dados com o auxílio de computadores cada vez mais poderosos, tanto em processamento quanto em capacidade de armazenamento. Diante disto, a extração de informação sem a ajuda de uma ferramenta torna-se uma tarefa árdua. Fayyad et al. (1996) propuseram um processo para a extração de conhecimento em banco de dados, denominado KDD – *Knowledge Discovery in Databases* – que consiste em "identificar padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados".

Uma das etapas desse processo, segundo Fayyad et al. (1996), é a mineração de dados que consiste da aplicação de algoritmos para a busca de informações nos dados por meio da verificação de relacionamentos entre seus atributos a fim de descobrir regras e/ou padrões significativos.

Os resultados obtidos da mineração de dados podem ser transcritos para um formato gráfico utilizando-se técnicas de visualização para que o usuário, conhecedor do domínio, tenha melhores condições de compreender as informações extraídas dos dados para a tomada de decisão. Desse modo, de acordo com pesquisas realizadas na literatura (Oliveira e Levkowitz, 2003; Shimabukuro, 2004; Rezende, 2005; Rabelo, 2007), técnicas de visualização podem ser usadas em conjunto com técnicas de mineração de dados.

No entanto, as técnicas de visualização podem ser aplicadas em qualquer fase do processo de descoberta de conhecimento, até mesmo sem o uso de um algoritmo de mineração de dados, devido a visualização ser capaz de representar as informações e conhecimentos extraídos sobre os dados originais de modo mais fiel, uma vez que os mesmos não passaram por nenhum processo adicional, tal como: transformação, seleção e integração.

Cabe ressaltar que a aplicação de uma técnica de visualização depende de vários fatores como, por exemplo, o tipo de dado a ser representado. Por esse motivo, é importante que sejam escolhidas técnicas de visualização adequadas, para que a visualização possa fornecer informações corretas a respeito dos dados representados graficamente.

Assim, este trabalho tem como objetivo definir diretrizes na escolha de técnicas de visualização a serem utilizadas como meio para a descoberta de conhecimento, considerando as características apresentadas pelos dados manipulados.

1.1. Contextualização

A necessidade de extrair conhecimento de grande quantidade de dados resultou no surgimento de métodos como a mineração de dados, comumente definida como um processo de extração de padrões e modelos (Mendonça e Sunderhaft, 1999). Alguns autores (Fayyad et al, 1996; Han e Kamber, 2001) definem que a mineração de dados faz parte de um processo maior, denominado KDD.

Os especialistas do domínio têm o papel de analisar e explicar as respostas geradas pela aplicação desse processo. Técnicas de visualização podem ser utilizadas para auxiliar na análise e interpretação dos resultados obtidos. A utilização de técnicas de visualização como subsídio no processo de extração de conhecimento sobre dados é mencionada em diversos campos de pesquisa, tais como: Análise Exploratória dos Dados; Mineração Visual de Dados; Visualização da Informação; Visualização do Conhecimento; Visualização de Dados; e Processo Analítico Visual (*Visual Analytics*). Cada uma dessas áreas é abordada em seções específicas deste trabalho, mas isso não significa que há uma fronteira formal estabelecida entre esses assuntos, pois uma mesma técnica de visualização pode ser empregada sob diferentes aspectos.

Na Análise Exploratória dos Dados são usados, geralmente, elementos gráficos comuns da estatística descritiva, como: histogramas, diagramas de caixa e diagramas de dispersão, com o objetivo de obter modelos matemáticos úteis sobre os dados analisados.

Grande parte das pesquisas inclusas na Visualização da Informação, segundo Oliveira e Levkowitz (2003), divide-se em duas categorias: Exploração Visual de Dados para a Mineração e Visualização de Modelos de Mineração, mas ainda existe um terceiro gênero, que é a Mineração Visual de Dados.

A Exploração Visual de Dados para a Mineração utiliza técnicas de visualização para subsidiar os métodos de extração do conhecimento, como o processo de mineração de dados. Tem por objetivo a detecção de padrões úteis e desconhecidos, realizando a análise por meio das representações gráficas dos dados, sem ter uma definição prévia de modelos ou hipóteses (Keim, 2002).

A Visualização de Modelos de Mineração consiste em visualizar os resultados de um algoritmo de mineração, permitindo maior entendimento sobre eles.

A Mineração Visual de Dados (*Visual Data Mining* - VDM) objetiva unir a visualização e o processo de análise de dados em uma única ferramenta, em vez de utilizar técnicas de visualização e mineração separadamente, resultando em um algoritmo visual de mineração.

As técnicas de visualização são importantes para explorar os dados e extrair conhecimento. Entretanto, Burkhard (2004) afirma que há uma dificuldade em aliar os métodos de visualização com a transmissão desse conhecimento adquirido entre as partes interessadas de uma instituição, constituindo um novo foco de pesquisa denominada Visualização do Conhecimento. Para Burkhard (2004), a Visualização do Conhecimento e a Visualização da Informação exploram a capacidade humana de processamento visual, mas diferem quanto ao objetivo: a Visualização da Informação utiliza a visualização como meio de explorar grande quantidade de dados e extrair novas percepções, ligada a um tipo de conhecimento que é baseado em fatos e informação; a Visualização do Conhecimento utiliza a visualização como meio de compartilhar ou disseminar conhecimento entre as pessoas, e integra vários campos de pesquisa, como a própria Visualização da Informação, Gestão do Conhecimento, Ciência da Comunicação e Arte Cognitiva.

Chen et al. (2008, p. 6), por sua vez, definem o termo Visualização de Dados (*Data Visualization*) pelo qual a visualização é utilizada para identificar a estrutura formada pelos dados, como também pode servir como meio de dedução da informação, em vez de apenas apresentá-la pronta. A ideia é aproximar os gráficos da modelagem estatística como meio para adquirir conhecimento, aliado ao poder computacional oferecido atualmente.

Outra área de estudo relacionada à visualização é o Processo Analítico Visual, denominada também pelo termo inglês *Visual Analytics* (Thomas e Cook, 2006; Keim et al., 2008), definida como “uma ciência que envolve o raciocínio analítico, subsidiada por interfaces visuais interativas” (Thomas e Cook, 2006; p.11).

Diante de todo este contexto, verifica-se que técnicas de visualização possuem um papel importante no processo de extração do conhecimento, independente da terminologia da área de estudo em questão. Assim, o foco deste trabalho está na representatividade das técnicas de visualização como ferramenta de descoberta de conhecimento sobre os dados.

1.2. Motivação

Diferentemente de Fayyad et al. (1996), que consideram a mineração de dados como uma etapa pertencente a um processo maior, o processo de KDD, há autores que consideram a mineração de dados, por si só, como um processo de extração de conhecimento novo, útil e não trivial.

Mendonça e Sunderhaft (1999), por exemplo, classificam a mineração de dados em três níveis de subprocessos, de acordo com as tarefas de mineração que executam, sendo consideradas seis tarefas: (1) Estimativa; (2) Classificação; (3) Associação; (4) Segmentação; (5) Visualização de dados e; (6) Exploração visual de dados. Estes três níveis estão correlacionados entre si e podem ser visualizados como uma pirâmide, ilustrada na Figura 1.1, a qual demonstra que a Exploração Visual Interativa é a atividade que provê fundamentação para os níveis subsequentes denominados Extração de Padrões e Construção de Modelos.

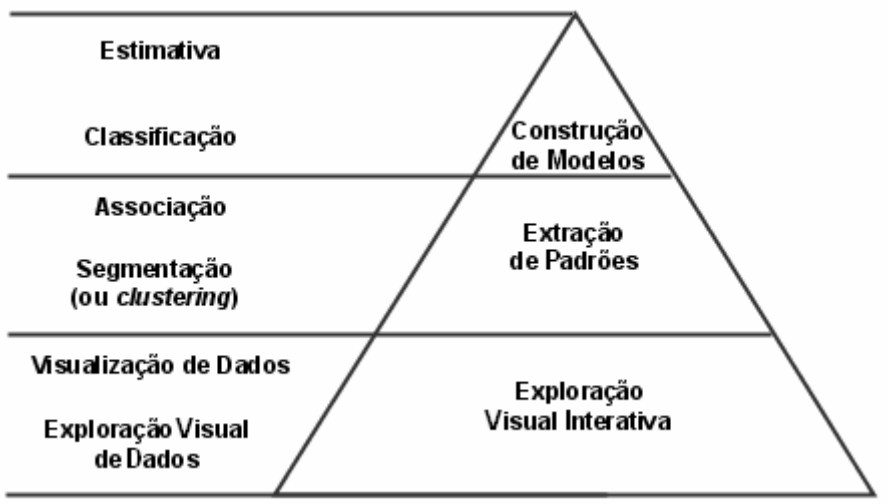

Figura 1.1: Níveis de mineração de dados (adaptado de Mendonça e Sunderhaft (1999))

De acordo com Mendonça e Sunderhaft (1999), o nível de Exploração Visual Interativa utiliza técnicas de visualização para transcrever os dados em formatos gráficos. Segundo estes autores, uma boa representação visual provê ao analista uma compreensão melhor sobre os dados. Por meio da tarefa de Exploração Visual de Dados, que permite ao analista interagir com as visualizações, é possível determinar como os dados estão correlacionados de forma a caracterizar o comportamento dos mesmos em uma determinada situação da aplicação. Isto promove hipóteses a serem examinadas no nível de Extração de Padrões, no qual os relacionamentos entre os dados são detectados de maneira automática, por meio de tarefas de mineração como Associação e Segmentação. No nível de Construção de Modelos, os padrões detectados são mapeados para um modelo com a utilização de tarefas de mineração como Estimativa e Classificação.

No entanto, a Exploração Visual de Dados aplicada sobre dados brutos ou originais pode ser suficiente para obter conhecimento satisfatório, uma vez que novos padrões podem ser detectados durante a exploração visual dos dados.

Porém, as características inerentes aos dados e as tarefas de mineração pretendidas devem ser analisadas ao aplicar técnicas de visualização de modo que as mesmas sejam adequadas para alcançar resultados corretos e satisfatórios. É neste contexto que concentra-se este trabalho.

1.3. Objetivos do trabalho

O objetivo principal deste trabalho é estabelecer diretrizes para escolha de técnicas de visualização a serem aplicadas na extração de conhecimento em banco de dados.

1.3.1. Objetivos específicos

Como objetivos específicos, têm-se:

- Identificar parâmetros que influenciam na escolha de técnicas de visualização;
- Selecionar técnicas de visualização para análise considerando os parâmetros identificados;
- Aplicar as técnicas de visualização selecionadas sobre bases de dados;
- Correlacionar as características existentes nos dados com as técnicas de visualização analisadas;
- Definir as diretrizes com base nos itens anteriores.

1.4. Metodologia da pesquisa

Sendo este presente trabalho classificado como uma pesquisa qualitativa, conforme a classificação dada por Silva e Menezes (2005), foi utilizada uma metodologia também qualitativa para conduzir o seu desenvolvimento, conhecida como Teoria Fundamentada em Dados (do inglês, *Grounded Theory*), uma metodologia investigativa para pesquisas qualitativas que visa a elaboração de uma teoria baseada em uma análise sistemática de um conjunto de dados, discutida com maiores detalhes no Capítulo 3.

A opção pelo uso desta metodologia deve-se à compatibilidade de seus princípios com o procedimento técnico adotado com o objetivo principal e com a característica exploratória deste trabalho. Segundo Silva e Menezes (2005), uma pesquisa com característica exploratória visa elucidar o problema por meio de levantamento bibliográfico, análise de exemplos que estimulem a compreensão, dentre outras atividades.

De acordo com essa metodologia, deve ser formada uma coleção de dados, denominada

amostragem teórica, sobre a qual serão realizadas análise e investigação. Neste trabalho, a principal fonte de coleta de dados concentra-se na bibliografia existente na literatura, correspondendo ao procedimento técnico baseado no levantamento bibliográfico mencionado por Silva e Menezes (2005, p.21). Neste tipo de procedimento, a pesquisa é "elaborada a partir de material já publicado, constituído principalmente de livros, artigos de periódicos e atualmente com material disponibilizado na Internet".

Por fim, a Teoria Fundamentada em Dados contribuiu para alcançar o objetivo principal do trabalho, servindo como base para o estabelecimento das diretrizes para a escolha de técnicas de visualização, por meio dos métodos de análise definidos por essa metodologia.

Dessa forma, o processo de desenvolvimento deste trabalho foi de acordo com as etapas descritas a seguir.

1) Revisão bibliográfica

   Por se tratar de um trabalho relacionado ao processo de extração de conhecimento, conceitos como mineração de dados e o processo de KDD fazem parte da revisão bibliográfica. Também foi realizado um levantamento sobre campos de estudos relacionados à visualização. Foram abordadas as áreas: Análise Exploratória dos Dados, Visualização da Informação, Mineração Visual de Dados, Visualização do Conhecimento, Visualização de Dados e Processo Analítico Visual.

2) Identificação de parâmetros que influenciam na escolha de técnicas de visualização

   Os trabalhos de: Keim e Kriegel (1996), que apresenta uma comparação entre técnicas de visualização; Dias (2002), que reuniu parâmetros relevantes para a escolha de técnica de mineração de dados; e Rabelo et al. (2008), que apresentam uma avaliação de técnicas de visualização para a tarefa de mineração; constituem a base referencial para o desenvolvimento do presente trabalho e os conceitos e critérios por eles desenvolvidos servem como modelo para o estabelecimento das diretrizes para a escolha de técnicas de visualização. Os parâmetros definidos nesta etapa do trabalho foram obtidos por meio dos métodos da Teoria Fundamentada em Dados.

3) Estudo de ferramentas para visualização

   Nesta etapa foram estudadas algumas ferramentas computacionais que podem ser utilizadas na geração dos tipos de visualização abordadas neste trabalho.

4) Seleção de técnicas de visualização e aplicação sobre bases de dados

   Foram utilizadas algumas das ferramentas estudadas na Etapa 3, que implementam

técnicas de visualização estudadas na Etapa 2. As representações gráficas geradas serviram para ilustrar o emprego de técnicas de visualização sobre bases de dados (aplicações práticas), sendo essas representações utilizadas como exemplo de identificação de quais características existentes nos dados podem influenciar direta ou indiretamente na escolha de uma determinada técnica. Esta etapa representa a triangulação dos dados, procedimento realizado para aumentar o acervo da amostragem teórica definida pela Teoria Fundamentada em Dados.

5) Definição de diretrizes que influenciam na escolha de técnicas de visualização
   Feito o levantamento dos parâmetros na Etapa 2 e, a partir dos resultados obtidos com a aplicação de técnicas de visualização na Etapa 3 e 4, foi possível, nesta etapa, elaborar as diretrizes que consistem da correlação entre as características identificadas nos dados e os fatores que influenciam na determinação de técnicas de visualização que podem ser utilizadas na extração do conhecimento. Tais diretrizes foram elaboradas como sendo o resultado final da aplicação da Teoria Fundamentada em Dados.

1.5. Organização do trabalho

Este trabalho de dissertação é composto por sete capítulos, cujos conteúdos são descritos a seguir.

Capítulo 1
Contém a introdução deste trabalho de dissertação, que inclui contextualização, motivação, objetivos, metodologia da pesquisa e organização do trabalho.

Capítulo 2
Traz a revisão bibliográfica que fundamenta a pesquisa desenvolvida, começando pela breve descrição de mecanismos de busca e extração de conhecimento em base de dados, como a mineração de dados e o processo de descoberta de conhecimento em banco de dados, comumente conhecido como processo de KDD. Esses métodos podem ser complementados por técnicas de visualização, abordadas por diferentes frentes de pesquisa, conforme apresentadas neste capítulo, que são: Análise Exploratória de Dados, Visualização da Informação, Mineração Visual de Dados, Visualização do Conhecimento, Visualização de Dados e Processo Analítico Visual.

Capítulo 3

Neste capítulo é descrita a metodologia aplicada na avaliação do presente trabalho, denominada Teoria Fundamentada em Dados, e são apresentadas as suas características e os procedimentos seguidos para que a condução de uma pesquisa qualitativa distancie-se da subjetividade do autor.

Capítulo 4

Neste capítulo são apresentados os parâmetros a serem considerados na escolha de técnicas de visualização, identificados durante a pesquisa, por meio do uso dos métodos da Teoria Fundamentada em Dados. De acordo com esses parâmetros são descritas características de algumas técnicas de visualização pertencentes a classes de: gráficos estatísticos, técnicas geométricas, técnicas iconográficas, técnicas orientadas a pixel e técnicas baseadas em grafos.

Capítulo 5

É descrito neste capítulo o procedimento de triangulação, pelo qual visualizações geradas por ferramentas computacionais constituíram uma segunda fonte de dados que foram analisadas conforme os parâmetros identificados no Capítulo 4. A triangulação serviu como um complemento para demonstrar a influência desses parâmetros na escolha de técnicas de visualização.

Capítulo 6

Tomando como base os resultados das análises dos parâmetros e sua associação com técnicas de visualização, realizadas nos capítulos anteriores, foi possível elaborar as diretrizes para a escolha de técnicas de visualização, que são descritas neste capítulo.

Capítulo 7

Contém as conclusões e sugestões para trabalhos futuros.

2. Revisão Bibliográfica

2.1. Considerações Iniciais

A busca por informações relevantes a partir de grande quantidade de dados culminou no desenvolvimento de técnicas como a mineração de dados que constitui uma fase do processo de descoberta de conhecimento (Fayyad et al., 1996).

A visualização pode ser utilizada para facilitar o entendimento sobre os resultados obtidos durante esse processo. Portanto, nos tópicos que seguem são abordados os conceitos básicos de mineração de dados e do processo de descoberta de conhecimento em banco de dados e, também, como as técnicas de visualização podem ser integradas a esse processo. Além disso, nesta seção são apresentadas, de forma geral, as áreas nas quais a visualização exerce papel principal no processo de extração de conhecimento. Assim, os conceitos anteriormente descritos formam a contextualização geral no qual se embasa este trabalho a respeito da utilização de técnicas de visualização.

2.2. Mineração de dados e o Processo de descoberta de conhecimento em banco de dados

O objetivo da mineração de dados é a extração de conhecimento implícito por meio da descoberta de padrões e regras significativas, a partir de grande quantidade de dados armazenados, de forma automática ou semi-automática, utilizando modelos computacionais construídos para descobrir novos fatos e relacionamentos entre dados, de forma repetida e interativa (Fayyad et al, 1996; Han e Kamber, 2001).

Para Fayyad et al. (1996) e Feldens et al. (1998), a mineração de dados constitui uma das etapas do processo de descoberta de conhecimento, conhecido também por KDD (*Knowledge Discovery in Databases*), as quais são definidas de diferentes modos por estes autores.

Fayyad et al. (1996), consideram o processo de KDD dividido em nove etapas, ilustradas na Figura 2.1, que compreendem: (1) Seleção: dados de interesse são escolhidos para o processo; (2) Criação de dados alvo: são focados atributos sobre os quais será executada a busca por conhecimento, formando um subconjunto de dados; (3) Pré-processamento dos dados: os dados selecionados passam por um processo de formatação e limpeza; (4) Transformação: são realizados processos de redução e projeção sobre os dados pré-processados; (5) Escolha de um modelo de mineração e ferramenta de mineração a ser utilizada; (6) Escolha dos algoritmos de mineração de dados; (7) Mineração de dados: são aplicados os algoritmos definidos na etapa 6 em busca de padrões para o modelo estabelecido na etapa 5; (8) Interpretação/avaliação dos resultados: compreensão dos padrões descobertos; (9) Utilização dos padrões descobertos: incorporação do conhecimento obtido ao processo de tomada de decisão.

Figura 2.1: *Processo de KDD segundo Fayyad et al. (1996)*

Feldens et al. (1998), por outro lado, consideram que este processo possui apenas três etapas, ilustradas na Figura 2.2: (1) Pré-processamento; (2) Mineração de dados; (3) Pós-processamento. O pré-processamento consiste das atividades que precedem a mineração de dados, que inclui análise, integração e transformação dos dados, podendo aproveitar as vantagens do uso de um *data warehouse*[1]. A etapa da mineração de dados compreende a aplicação de algoritmos de mineração, que podem ser reaplicados diversas vezes para calibrar seus parâmetros a fim de obter maior eficiência. O pós-processamento pode ser definido pelas operações de filtragem, estruturação e ordenação dos resultados da mineração para então serem apresentados ao usuário.

[1] Um *data warehouse* é uma coleção de dados orientada por assuntos, integrada, variante no tempo e não volátil, que tem por objetivo apoiar aos processos de tomada de decisão (Inmon, 1997).

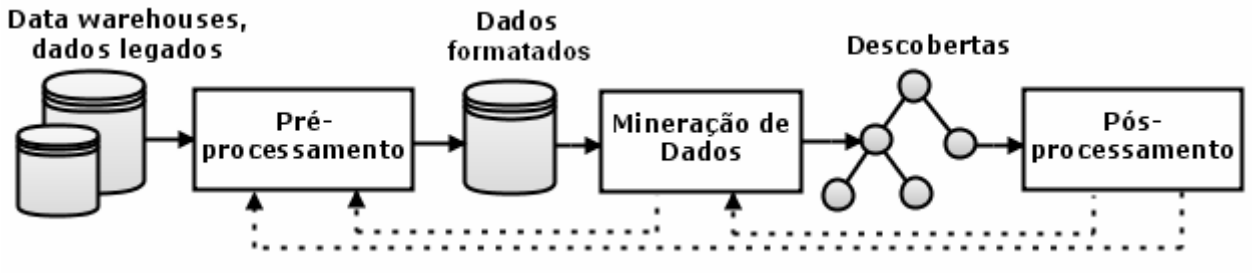

Figura 2.2: Processo de KDD segundo Feldens et al. (1998)

De modo geral, primeiramente é feita a escolha das fontes de dados a serem utilizadas e a definição dos objetivos. Uma parcela desses dados é selecionada, pré-processada e submetida a métodos e ferramentas (tarefas e algoritmos de mineração de dados) adequados com o objetivo de encontrar padrões e/ou modelos que representem o conhecimento obtido. Depois de extraídos, os padrões são pós-processados e o conhecimento adquirido é avaliado quanto a sua qualidade e utilidade para determinar a viabilidade de sua utilização no apoio a algum processo de tomada de decisão. É nessa etapa que a visualização serve de instrumento para aumentar o entendimento sobre os resultados desse processo.

Em ambas as versões apresentadas por Fayyad et al. (1996) e Feldens et al. (1998), cada fase definida pode ser repetida em várias iterações. Além de iterativo, KDD também é um processo interativo por exigir a presença de um especialista do domínio para interagir com o processo, refinando assim o conhecimento.

2.2.1. Integração da visualização com o processo de descoberta de conhecimento

Segundo Ankerst (2001) e Keim (2002), a etapa da mineração de dados é mais eficaz quando há a coordenação de um especialista durante o processo de exploração de dados. A integração de técnicas de visualização com técnicas de mineração de dados agrega qualidade e agilidade a esse processo. Neste caso, a visualização é um agente facilitador para o entendimento dos resultados obtidos.

Ankerst (2001) distingue a integração da visualização com a exploração de dados em três grupos, como mostra a Figura 2.3. No primeiro, Visualização Anterior (Figura 2.3.a), as técnicas de visualização são aplicadas antes do algoritmo de mineração para identificar possíveis problemas, realizando assim a pré-visualização dos dados. No segundo grupo, Visualização Posterior (Figura 2.3.b), a visualização é aplicada sobre o resultado dos algoritmos de mineração,

a fim de melhorar o entendimento acerca dos padrões descobertos. No terceiro grupo, Visualização Fortemente Integrada (Figura 2.3.c), a visualização é aplicada concomitante à execução do algoritmo de mineração, sendo possível a visualização de seus resultados intermediários. Neste último grupo, o usuário pode avaliar melhor todo o processo de descoberta de conhecimento.

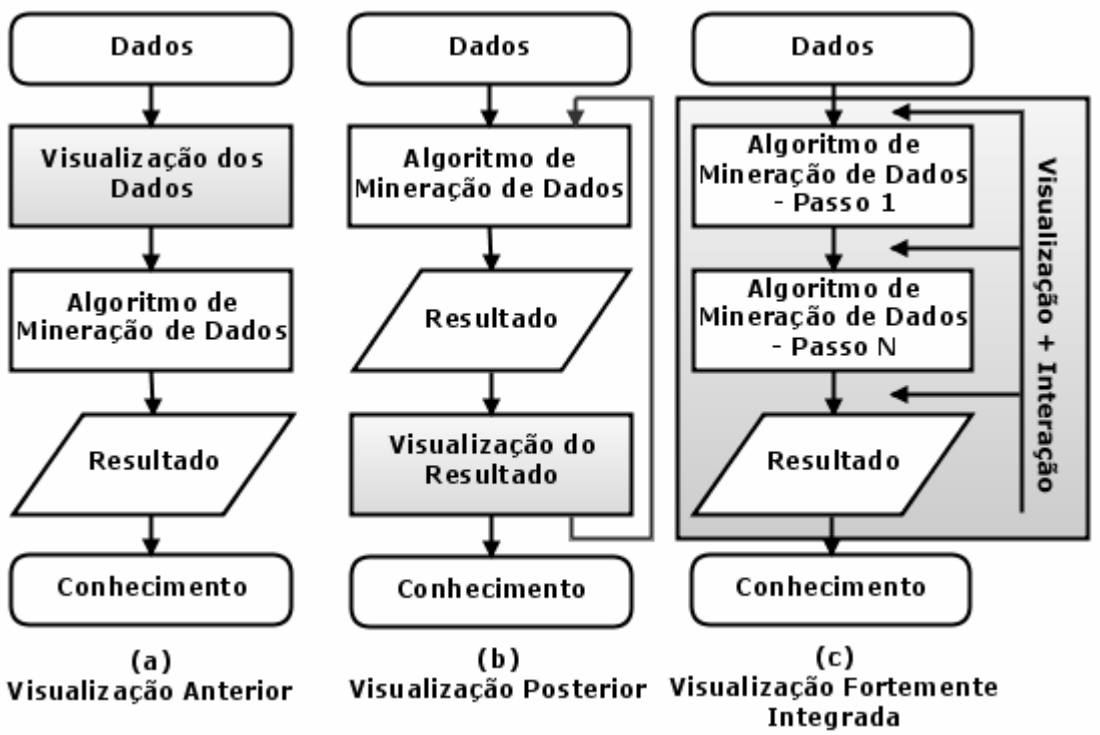

Figura 2.3: Três modos de integração das técnicas de visualização com a exploração de dados (adaptado de Ankerst, 2001 apud Neto, 2008)

2.3. Visualização

De acordo com Shimabukuro (2004) e Nascimento e Ferreira (2005), dentre os sentidos de percepção humana, a visão é a que possui maior capacidade de conceber informações do mundo, por ser rápida e paralela, sendo possível focar a atenção em um ponto de um cenário visual e ainda perceber o que acontece ao redor. Além disso, o sistema de visão humano possui aptidão para a percepção de padrões que permitem a interpretação de comportamentos, tendências, relacionamentos e exceções existentes nos dados. Keim (2002) afirma que a visualização facilita a exploração de dados não homogêneos e com inconsistências (ruídos) de modo intuitivo, por não exigir um conhecimento profundo em matemática ou estatística.

Dessa forma, a visualização explora as vantagens da percepção da visão para auxiliar na

compreensão de dados e informações a fim de apoiar a cognição humana ao associar um conjunto de dados a uma representação gráfica, facilitando o entendimento e possibilitando a descobertas de novos conhecimentos (Alexandre e Tavares, 2007).

Assim, geralmente é mais fácil compreender uma visualização gráfica do que um amontoado de dados dispostos em uma tabela. Segundo Keim (2002), a representação gráfica de dados é um modo mais rápido para prover ou transmitir conhecimento e, frequentemente, apresenta melhores resultados. A visualização serve como um veículo para a transmissão do conhecimento em diversas áreas. Este recurso tem sido utilizado desde longa data, como mostra os trabalhos de Friendly (2008) e Friendly e Denis (2008), discutidos a seguir.

2.3.1. Histórico da visualização

Desde os tempos antigos, datados por volta do início do século XVII, a representação visual de dados é utilizada para transmitir ou documentar o conhecimento. Friendly (2008) enumera diversos exemplos de representações gráficas que reconstroem o histórico da visualização de dados, desde os primeiros esboços gráficos da Cartografia e Astronomia, dos quais as técnicas de construção de gráficos expandiram-se e evoluíram para as mais diversas áreas. Hoje tem-se a visualização de dados aplicada na área médica, geográfica, empresarial, dentre outros, bem como no desenvolvimento de diversas técnicas de visualização de informação.

Nesta seção é descrita uma breve cronologia da evolução da visualização de dados, baseado nos trabalhos de Friendly (2008) e Friendly e Denis (2008). Observando a Figura 2.4, pode-se ver que a linha de evolução dos gráficos inicia-se por volta do século XVI, com os primeiros mapas astronômicos e diagramas geométricos. Um exemplo de gráfico produzido nesta época está ilustrado na Figura 2.5.

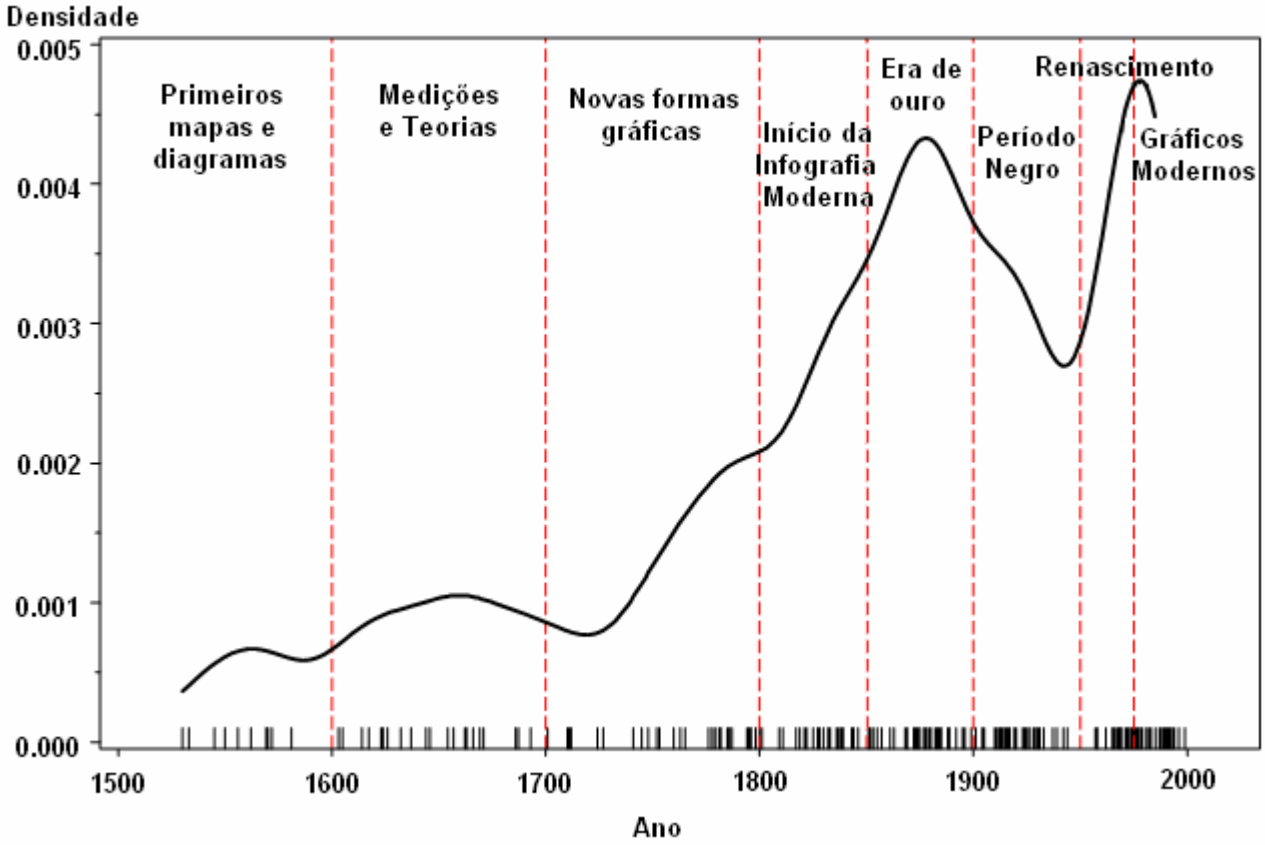

Figura 2.4: Distribuição dos marcos do histórico da visualização. O gráfico mostra a densidade estimada para a distribuição de 248 marcos históricos identificados desde o ano 1500 até o presente (adaptado de Friendly, 2008)

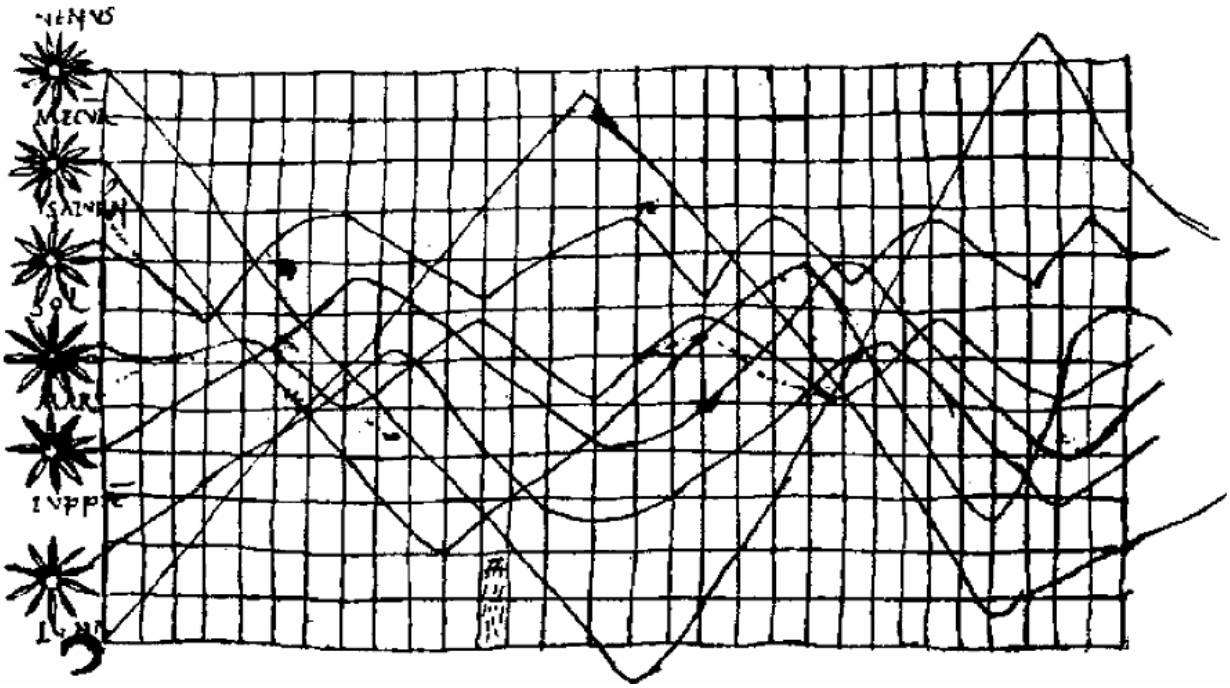

Figura 2.5: *Inclinação das órbitas planetárias ao longo do tempo – ano 950 : um dos primeiros gráficos a mostrar variáveis, que descreve a movimentação de corpos celestiais; o eixo vertical representa a inclinação das órbitas planetárias, e o eixo horizontal mostra tempo, dividido em 30 intervalos (Friendly e Denis, 2008)*

Ao longo do século XVII, instrumentos para medições de grandezas físicas (distância, tempo, espaço) foram desenvolvidos ao mesmo tempo em que os estudos sobre estatísticas foram avançando. Este período contribuiu com os primeiros passos para o desenvolvimento dos gráficos estatísticos que conhecemos hoje. Acredita-se que o gráfico feito por Michael Florent vanLangren[2] em 1644, ilustrado na Figura 2.6, seja a primeira representação visual de dados estatísticos (Tufte, 1997; p.15 apud Friendly, 2008).

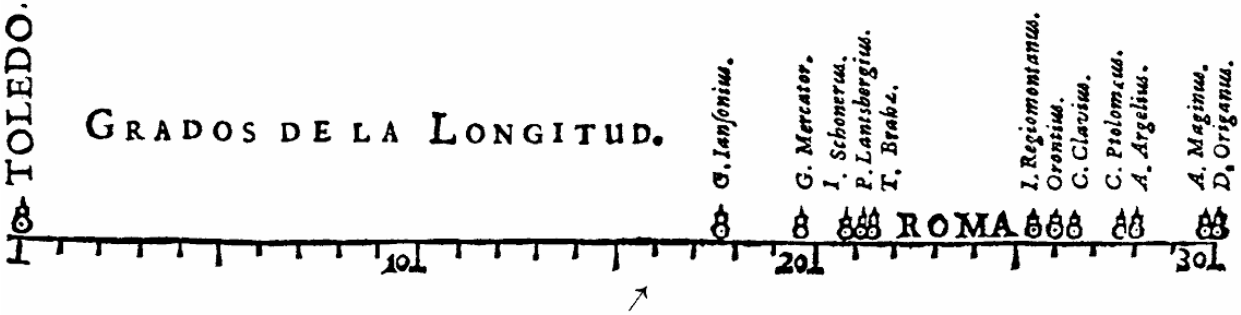

Figura 2.6: Gráfico de vanLangren, de 1644, que mostra as 12 referências de longitude de Toledo a Roma calculados por famosos astrônomos da época (Mercator, Tycho Brahe, Ptolemeu, entre outros). A verdadeira medida é de 16º30', apontada pela seta, demonstrando que as estimativas anteriores ultrapassaram a medida real (Friendly e Denis, 2008)

Continuando no século XVIII, Friendly e Denis (2008) citam William Playfair (1759-1823), economista inglês, conhecido por introduzir os gráficos estatísticos que são amplamente utilizados até os dias atuais. Este economista publicou, em 1786, os primeiros gráficos de barras e histogramas e, em 1801, apresentou os gráficos em pizza e gráficos circulares.

Friendly (2008) denomina como "Era de ouro dos Gráficos Estatísticos" o período próximo a 1850, no qual ocorreram várias inovações gráficas. Assim, no século XIX, todas as formas de gráficos estatísticos conhecidos hoje foram desenvolvidas nesta época.

Entretanto, de acordo com Friendly (2008), no início do século seguinte, ocorreram poucas inovações gráficas, sendo esta época denominada por "*Período Negro dos Gráficos da Estatística*", como indica o gráfico na Figura 2.4. Somente a partir da década de 80 do século XX que a visualização retoma forças com o advento dos computadores. Neste período, denominado como "*Renascimento da Visualização*", John W. Tukey foi um dos primeiros estudiosos a analisar dados por meio dos gráficos, fato que começou a ganhar importância com o advento de softwares para a manipulação de dados e surgimento de novos métodos de representação gráfica impulsionando o crescimento da área de Visualização. Assim, a evolução

[2] Michael Florent vanLangren (1600 - 1675): Astrônomo belga que serviu à corte espanhola.

dos computadores possibilitou a evolução de técnicas de visualização de dados e, atualmente, existem vários métodos modernos de representação gráfica, discutidos nas próximas seções. Embora existam softwares capazes de gerar gráficos automaticamente, é necessário o conhecimento sobre boas práticas de construção de gráficos. Unwin (2008) traz uma abordagem nesse sentido, conforme apresentado na seção a seguir.

2.3.2. A escolha de uma boa representação gráfica

Unwin (2008) discursa sobre a construção de gráficos e os cuidados que deve-se ter ao fazê-lo, de modo que o gráfico possa de fato alcançar o objetivo de transmitir informações ao leitor, da forma que o autor deseja e de maneira clara e correta. Primeiramente, o gráfico deve estar inserido em um contexto, servindo como complemento para uma discussão relatada em um trabalho. Um bom gráfico tem de ter uma boa construção em relação à apresentação do conteúdo e ao contexto. Para isto, existem alguns princípios que devem ser seguidos no momento da construção do gráfico, caso contrário, um gráfico pode causar mais confusão do que complementar o entendimento sobre determinado assunto. Escalas mal desenhadas; emprego de visualização 3D ou 2D para representar áreas aparentemente proporcionais aos valores que na realidade não são; informações comprimidas em um espaço muito pequeno; são exemplos de gráficos pouco inteligíveis.

A escolha de uma representação gráfica depende de vários fatores, como o propósito do gráfico e o tipo dos dados a ser representado (por exemplo, gráficos de pizza são adequados para exibir partes de um pequeno grupo de categorias). Várias representações podem ser empregadas para um mesmo conjunto de dados, de acordo com as características dos dados que deseja-se enfatizar.

Os softwares que são atualmente desenvolvidos para representação gráfica são capazes de controlar as visualizações de maneira intuitiva, constroem gráficos padronizados, evitando erros grosseiros, tornando os dados mais legíveis para quem está familiarizado com eles. Por outro lado, Unwin (2008) afirma que esses softwares, por mais sofisticados que sejam, não garantem que os gráficos construídos sejam de fato bons gráficos, sendo a qualidade da visualização dos dados dependente do bom uso de técnicas adequadas ao conteúdo dos dados.

Diante deste fato, Unwin (2008) apresenta uma discussão sobre os fatores que influenciam na leitura e construção de um gráfico, como: (a) escala; (b) ordenação e organização dos valores que são desenhados no gráfico; (c) os papéis da legenda**;** (d) como deve estar a posição do gráfico em relação ao texto**;** (e) o tamanho, a estrutura e a proporção do gráfico; e

ainda, (f) como a cor pode influenciar na leitura de um gráfico. Apesar de existirem softwares que geram os gráficos, nem sempre o resultado é satisfatório. Por isso, convém considerar esses fatores, descritos a seguir, ao desenhar um gráfico.

a) Escala

A escala é importante para a organização de dados categóricos. A definição da escala torna-se mais difícil quando os dados são contínuos. Uma boa prática é estender a escala para valores além dos limites observados e utilizar valores arredondados e múltiplos de uma constante nos eixos. Não é obrigatório ter o valor zero na escala como o ponto base, embora isso seja o mais comum. O significado dos dados e o modo que foram coletados devem ser considerados na decisão da escala. Outro ponto importante é a definição de rótulos nos eixos: muitos nomes podem causar confusão, poucos podem dificultar a avaliação dos valores. A Figura 2.7 exibe dois gráficos para o mesmo conjunto de dados, mas com escalas diferentes.

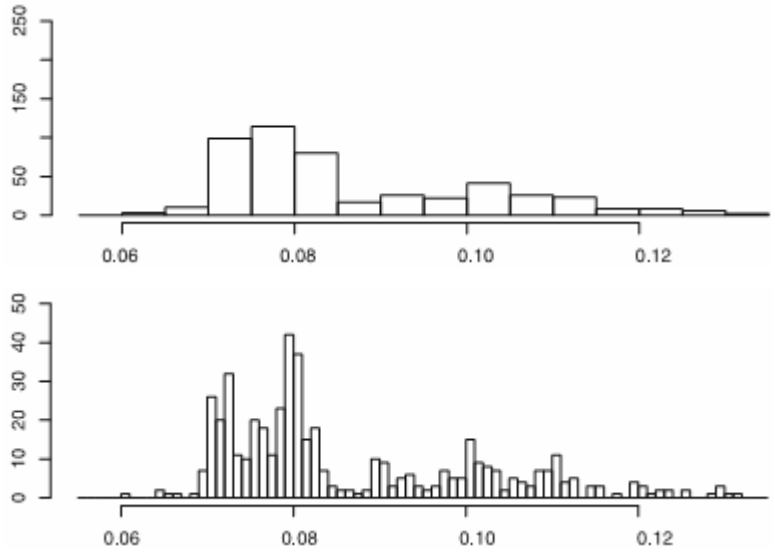

Figura 2.7: Exemplo de gráfico com escalas (Unwin, 2008)

b) Ordenação e organização

A posição e a ordem em que os atributos dos dados são posicionados no gráfico influenciam no resultado final da visualização. Exemplos de técnicas que requerem este cuidado são as Coordenadas Paralelas e o Gráfico em Mosaico. Quando o atributo não possui uma ordem natural, recomenda-se que a ordenação seja de acordo com algum tipo de categoria. As categorias podem ser ordenadas por tamanho ou por um segundo atributo.

c) Legenda

A legenda em forma de texto explicativo deve acompanhar e explicar o gráfico de modo completo, conciso e claro, informando inclusive a fonte dos dados. O gráfico por si só deve ser objetivo ao ilustrar os dados, sendo a função da legenda apenas sintetizar a informação. Um texto explicativo muito extenso pode ser um indício de que há muita informação além do que o gráfico deveria conter. Recomenda-se que a legenda que descreve símbolos e cores utilizados para diferenciar um conjunto de dados deve estar inclusa no desenho do gráfico, para evitar que fique em uma região separada, dificultando a sua leitura.

d) Posição

O gráfico ilustra uma ideia exposta pelo texto. Por isso, é conveniente posicioná-lo na mesma página por razões práticas durante a leitura, embora isso nem sempre seja possível.

e) Tamanho, Estrutura e Proporção

O tamanho de um gráfico deve ter medidas de modo que as informações dispostas nele apareçam claramente. Molduras podem ser desenhadas em torno do gráfico para fazer uma separação do texto ou de outro gráfico. A proporção entre a altura e a largura da imagem é significativa para a compreensão do gráfico, como mostra a Figura 2.8: os dados são os mesmos para os três gráficos, mas possuem altura e largura em proporções diferentes. Os dados representam os tempos de 80 corredores da última volta *vs.* a primeira em uma corrida de 100 Km. A linha traçada indica uma relação linear entre os corredores mais rápidos e os mais lentos.

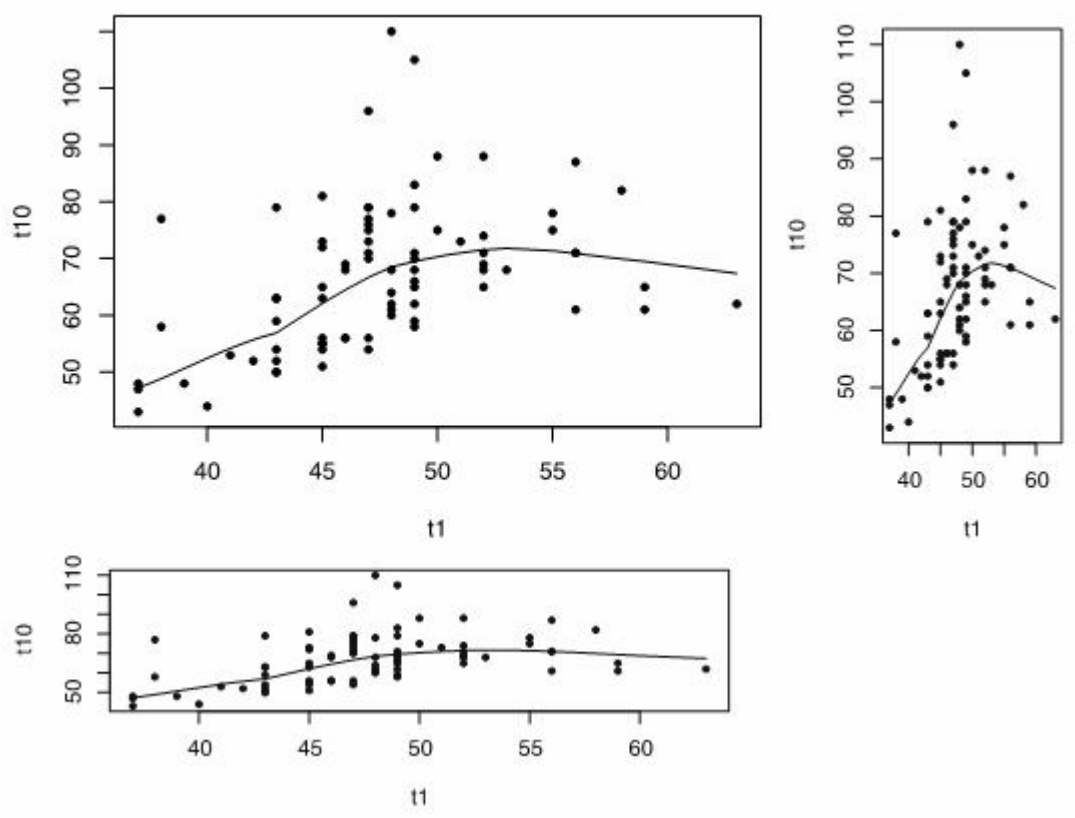

Figura 2.8: Gráficos que representam o mesmo conjunto de dados, com diferentes proporções de altura e largura (adaptado de Unwin, 2008).

f) Cores

A utilização de cores no gráfico pode ser um meio eficaz para a construção de gráficos como, por exemplo, utilizar tonalidades diferentes de cores para indicar valores de uma escala numérica (Green, 1998). Entretanto, alguns cuidados devem ser tomados para que isto seja feito de maneira correta: deve-se considerar que algumas pessoas não conseguem enxergar determinadas cores; algumas carregam um significado implícito, como o vermelho, que geralmente é associado com “perigo” ou “prejuízo”; além de sua utilização obedecer a um gosto pessoal.

2.3.3. Campos de Pesquisa sobre Visualização

A representação visual de dados é tema de estudo de vários campos de pesquisa, dentre os quais são descritos nesta seção os seguintes: (a) Análise Exploratória de Dados; (b) Visualização da Informação; (c) Mineração Visual de Dados; (d) Visualização do Conhecimento; (e) Visualização de Dados; (f) Processo Analítico Visual. Cabe ressaltar que não há uma delimitação clara entre esses campos de estudo, sendo o objetivo desta seção apenas levantar as peculiaridades de cada área.

Assim, inicialmente, é apresentada a Análise Exploratória de Dados que utiliza gráficos estatísticos durante o processo de análise de dados. Na sequência, é descrita a área Visualização da Informação, na qual são utilizadas técnicas de visualização mais sofisticadas em comparação com as utilizadas pela Análise Exploratória dos Dados. Logo após, é feita uma abordagem sobre Mineração Visual de Dados, cujo princípio é acoplar técnicas de visualização a técnicas de mineração. Também é exposta a ideia trazida pela Visualização do Conhecimento que foca o compartilhamento e a disseminação do conhecimento pela visualização. Em seguida, é apresentada a Visualização de Dados, área para qual o papel da representação gráfica é distinto em duas funções: apresentação ou exploração de dados. Por fim, é discutida a ideia do Processo Analítico Visual (*Visual Analytics*), uma área recente que coloca a visualização como mediadora central no processo de extração do conhecimento, em que não há etapas formalmente definidas, sendo baseada na forte interação do analista com o ambiente de análise de dados, composto por técnicas e ferramentas, abrangendo visualização, métodos de mineração de dados, análises estatísticas, percepção e cognição, gerenciamento de dados, entre outros, formando uma área multidisciplinar, bem como as outras áreas apresentadas.

a) Análise Exploratória de Dados

Análise Exploratória de Dados (AED) – do inglês *Exploratory Data Analysis* – é um termo cunhado em 1977 por John W. Tukey, pesquisador que prestou contribuição importante ao período do “renascimento da visualização de dados”, como apresentado no trabalho de Friendly (2008) visto na seção 2.3.1. O princípio da AED é analisar dados em busca de informações potencialmente úteis sem haver um modelo ou hipótese pré-concebida sobre eles, propondo uma abordagem para explorar os dados de modo que isto revele sua estrutura ou modelo. O procedimento seguido durante a análise dos dados depende do quanto sabe-se sobre eles para aplicar uma técnica de visualização adequada. O analista de dados é responsável pela condução desse processo e pelas escolhas das técnicas apropriadas (Eick e Wills, 1995; Morgenthaler,

2009).

A maioria das técnicas de AED é formada por gráficos estatísticos por meio dos quais podem ser detectados padrões e comportamentos presentes nos dados (NIST/SEMATECH, 2006). Além disso, a representação visual dos dados por meio desses gráficos possibilita inferências estatísticas e derivações de modelos. As técnicas de visualização podem ser combinadas entre si, bem como com outras técnicas da estatística para realizar a análise dos dados. Alguns exemplos de gráficos utilizados na AED são apresentados na Tabela 2.1, os quais também são descritos com mais detalhes no Capítulo 4.

Tabela 2.1: Exemplos de gráficos utilizados na análise exploratória de dados (NIST/SEMATECH, 2006)

Gráfico	**Função**
Histograma	Sintetizar a distribuição de um conjunto de valores de um atributo
Diagrama de dispersão (*scatter plot*)	Verificar relacionamento entre dois atributos
Diagrama de caixas (*box plot*)	Demonstrar a localização e variação entre diferentes grupos de dados e determinar se um fator possui efeito significante conforme a localização ou variação
Diagrama de estrelas (*star plot* ou *star glyph*)	Permitir a exibição de dados com múltiplos atributos. Cada parte da estrela representa a observação para um atributo.

De acordo com Myatt (2007) e (NIST/SEMATECH, 2006), a abordagem adotada pela AED segue os seguintes passos:

Problema → Dados → Análises → Modelo → Conclusões

Cada um destes passos está descrito a seguir.

1. Definição do problema – Consiste em delinear o plano de análise para encontrar as soluções para o problema a ser resolvido. Os artefatos gerados pela exploração de dados devem estar claramente definidos.

2. Preparação de dados – Envolve a seleção e preparação dos dados que serão analisados, incluindo limpeza e transformação para um formato apropriado ao processo de análise.

3. Implementação da análise – Compõe-se de três tarefas principais: sintetização dos dados; busca por relacionamentos implícitos; e elaboração de predições. A primeira constitui um processo que organiza e reduz os dados para interpretação, mas sem anular qualquer informação importante. A segunda baseia-se na exploração dos dados a partir de vários ângulos a fim de detectar padrões, comportamentos ou tendências ainda desconhecidas. A última refere-se a prognosticar eventos com base nas informações obtidas das tarefas

anteriores. A Figura 2.9 mostra que essas três tarefas estão interligadas e utilizam algumas técnicas em comum descritas em sequência.

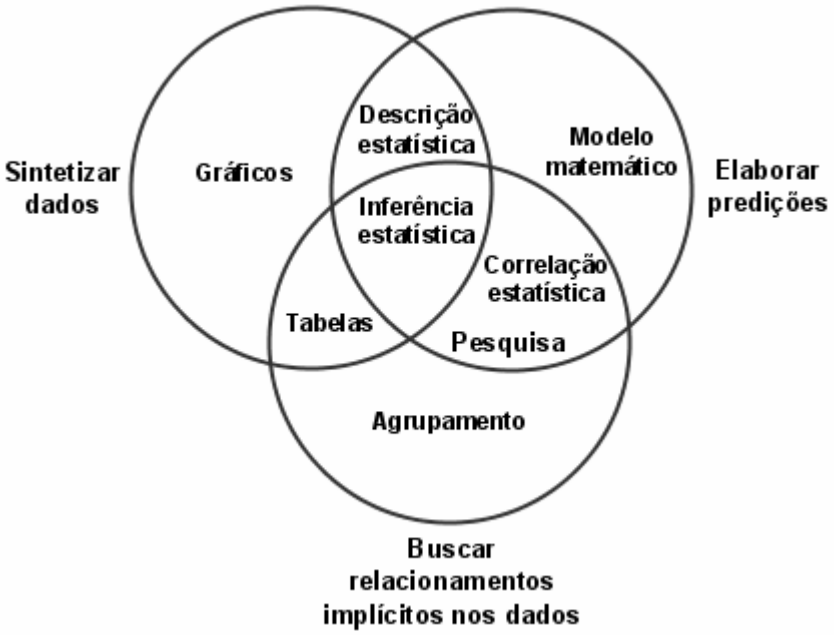

Figura 2.9: Tarefas e métodos da análise de dados (adaptado de Myatt (2007))

- Tabelas – Os dados brutos podem ser sintetizados e organizados de vários modos em tabelas.
- Gráficos – A representação gráfica dos dados facilita a identificação de padrões e relacionamentos.
- Descrições estatísticas – Formas de resumir informação sobre determinado dado, como um valor de média ou valores extremos sobre o conjunto total de elementos.
- Inferências estatísticas – Métodos que permitem que determinada afirmação sobre os dados seja feita com certo grau de confiança.
- Correlações estatísticas – Técnicas estatísticas que quantificam os relacionamentos presentes nos dados.
- Pesquisa – Consultas específicas a respeito dos dados podem ser úteis quando se quer compreender ou formular alguma conclusão com outras informações.
- Agrupamento – Métodos para agrupar os dados em grupos menores para responder potencialmente às questões.
- Modelos matemáticos – Equação matemática ou processos capazes de formar predições.

4. Apresentação dos resultados – Os resultados obtidos são apresentados às partes interessadas e podem servir de informação a outras tarefas ou serem incorporadas a um repositório de informações como planilhas eletrônicas ou páginas da *web*.

b) Visualização da Informação

A exploração de dados em busca de informações úteis torna-se uma tarefa mais custosa à medida que o volume de dados aumenta. Para facilitar esta atividade, técnicas de visualização permitem que grande quantidade de dados seja representada graficamente, tornando mais simples a detecção de padrões, comportamentos e tendências que possam estar contidos nos dados, que são fatores relevantes no momento de tomar decisões (Shneiderman, 1996).

Segundo Oliveira e Levkowitz (2003), a classificação das técnicas de visualização pode ser um guia na escolha de uma dessas técnicas de visualização, pois uma classe apresenta as características comuns sobre os dados representados. Entretanto, isso não significa que a escolha deve limitar-se a um determinado tipo de classificação existente, sendo possível que diferentes técnicas possam ser aplicadas conforme o objetivo da busca de conhecimento. Nesta seção são apresentadas as taxonomias clássicas das técnicas de visualização, de acordo com dois autores, (I) Shneiderman (1996) e (II) Keim (2002).

I) Classificação das técnicas de visualização segundo Shneiderman

Shneiderman (1996) define princípios fundamentais para a busca visual da informação, conhecidos como *Visual Information Seeking Mantra*: *overview first, zoom and filter, then details-on-demand*. Isto significa basicamente: visualizar todos os dados de modo geral, focar e filtrar, para então buscar detalhes conforme necessário. Assim, o usuário primeiro deve visualizar os dados completamente para identificar os padrões e escolher um ou dois que sejam de seu interesse. Para analisar os padrões, o usuário deve ter possibilidades de interagir sobre a visualização, explorando um subconjunto de dados em busca de informações, buscando por mais detalhes e descartando itens que sejam irrelevantes.

Seguindo estes critérios, Shneiderman (1996) classifica as técnicas de visualização de acordo com as tarefas a serem realizadas pelo usuário e pelos tipos de dados que são manipulados, sendo sete as tarefas enumeradas:

- Visão geral: adquire uma visão geral sobre todos os dados;
- Zoom: foca os itens de interesse;

- Filtragem: descarta os itens que não interessam;
- Detalhar sob demanda (*details-on-demand*): seleção de um item ou um grupo e obtenção de mais detalhes quando for necessário;
- Relacionar: visualiza relacionamentos entre os itens;
- Histórico: mantém o histórico das ações para tornar possíveis as opções de desfazer, refazer e refinar progressivamente;
- Extração: permite a extração de subconjuntos e parâmetros de consulta.

Shneiderman (1996) ainda classifica as técnicas de visualização pelos tipos de dados, que podem ser unidimensionais, bidimensionais (2D), tridimensionais (3D), temporais, multidimensionais, hierárquicos e grafos. Os dados unidimensionais podem representar documentos de texto, códigos-fonte de programas, listas de nomes que podem ser organizadas de alguma maneira. Os dados bidimensionais englobam dados de mapas geográficos. Dados tridimensionais representam objetos do mundo real (como o corpo humano, moléculas, construções de edifícios, fenômenos naturais) e suas relações com outros objetos. Segundo este autor, os dados tridimensionais são objetos de estudo no campo de Visualização Científica, que especifica-se na visualização desse tipo de dados (Shneiderman, 1996; Rhyne, 2003). Os dados temporais são registros que possuem algum significado de tempo. Dados multidimensionais representam as tabelas do banco de dados que contêm registros com vários atributos. Dados hierárquicos representam um conjunto de dados em que cada elemento possui uma ligação com um elemento-pai, exceto a raiz da árvore de hierarquia; dados que não enquadram-se em uma hierarquia, mas mantém um relacionamento entre os elementos, podem ser representados por grafos.

II) Classificação das técnicas de visualização segundo Keim

Keim (2002) considera três critérios para a classificação das técnicas de visualização: (1) tipo de dados a ser visualizado; (2) técnicas de visualização e (3) técnicas de interação e distorção utilizadas.

Os dados são distintos em unidimensionais (1D), bidimensionais (2D), multidimensionais (nD), texto e hipertexto, hierarquias/grafos e algoritmos/softwares. Quanto às técnicas de visualização, estão classificadas em gráficos convencionais (para representação de um a três atributos), técnicas geométricas, técnicas iconográficas, técnicas orientadas a pixel e técnicas baseadas em dimensões. Técnicas de interação e distorção são mecanismos para interagir com a representação gráfica, que podem ser Projeção, Filtragem, Zoom, Distorção e Ligação & Seleção

(*Link & Brush*). A taxonomia das técnicas de visualização estabelecida por Keim (2002) segundo esses critérios é apresentada a seguir.

1) Classificação pela natureza dos dados

Como exemplos de dados unidimensionais, têm-se os dados temporais, para os quais cada marco de tempo pode ser associado um ou vários valores de dados. Shneiderman (1996) faz a distinção entre tipos de dados unidimensionais e temporais, sendo que os dados temporais apresentam pontos de tempo inicial e final e os registros podem ser sobrepostos. Assim, as tarefas associadas a esse tipo de dado podem ser buscas de eventos anteriores, posteriores ou durante determinado período de tempo.

Os dados bidimensionais servem para representar dados geográficos, cujas dimensões são a latitude e longitude que podem ser visualizados por meio de gráficos com coordenadas *x* e *y*.

Os dados multidimensionais são aqueles encontrados em banco de dados e possuem mais de três atributos, sendo ineficaz a utilização de gráficos 2D ou 3D. Para a visualização desse tipo de dados, técnicas como Coordenadas Paralelas podem ser utilizadas. Keim (2002) faz a distinção de dados que referem-se a textos ou páginas da *web*, que não podem ser descritos em função de dimensão. Assim, uma transformação sobre esses dados deve ser feita para que posteriormente seja aplicada uma técnica de visualização (Havre et al., 2002; Starre e Vries, 2005; Mao et al., 2007).

Os dados hierárquicos e os grafos representam os relacionamentos que existem entre os elementos do conjunto de dados (Herman et al., 2000).

Outra distinção da classificação de Keim (2002) para a de Shneiderman (1996) é quanto à classe de dados algoritmos e softwares. O segundo autor classifica os códigos de programas como dados unidimensionais, enquanto que o primeiro considera que a visualização dessa categoria de dados tem como objetivo auxiliar no desenvolvimento do software, no entendimento do fluxo de dados pelos algoritmos, compreensão das linhas do código fonte por meio de gráficos como modo de visualizar erros ou apoio à depuração do código (Zeckzer et al., 2008; Sensalire et al., 2008).

2) Classificação pelas técnicas de visualização

Keim (2002) distingue as técnicas de visualização em cinco grupos: (1) gráficos 1D-3D; (2) técnicas baseadas em ícones; (3) técnicas geométricas ou geometricamente transformadas; (4) técnicas orientadas a pixel e (5) técnicas baseadas em dimensões. A seguir é detalhada cada uma dessas classes.

- Gráficos 1D-3D – correspondem aos gráficos amplamente utilizados na estatística (gráficos de barras, gráficos de pizza, gráficos de dispersão, etc.).
- Técnicas iconográficas – a visualização é feita mapeando os atributos de um dado para as propriedades de um ícone, ou glifo, que variam conforme os valores dos atributos. Um exemplo típico são as faces de Chernoff (Figura 2.10), para as quais é realizado o mapeamento entre os atributos dos dados e as propriedades da face – o formato do nariz, boca, olhos, rosto são alterados conforme o mapeamento.

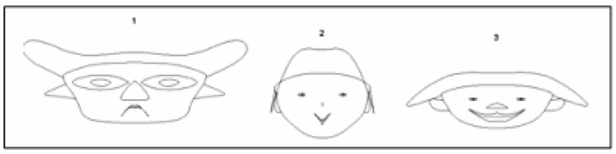

Figura 2.10: *Exemplo de Faces de Chernoff (Rabelo, 2007)*

- Técnicas geométricas ou geometricamente transformadas – são feitas transformações e projeções geométricas sobre um conjunto de dados multidimensionais com o objetivo de produzir visualizações que forneçam informações úteis. Como exemplo, a matriz de dispersão (*scatterplots*), matriz de dispersão 3D e coordenadas paralelas (Figura 2.11).

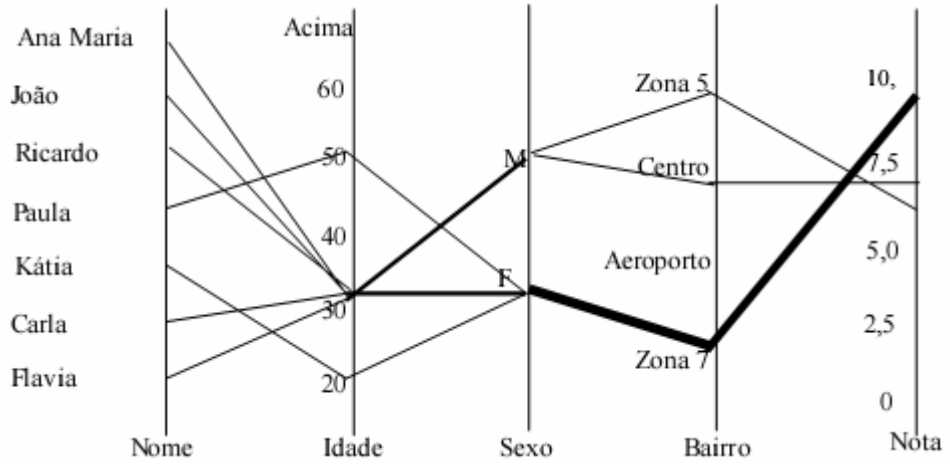

Figura 2.11: Exemplo de Coordenadas Paralelas com dados fictícios. Pela visualização pode-se notar que a maioria das pessoas relacionadas tem aproximadamente 30 anos, sendo a maior parte delas do sexo feminino, residentes na zona 7 e obtiveram notas próximas de 10. (Rabelo, 2007)

- Técnicas orientadas a pixel[3] – cada atributo é mapeado para uma cor de pixel, e distribuídos sobre a tela de exibição dividida em *n* janelas, cada qual correspondente a

[3] Aglutinação do termo em inglês ***Pic**ture **El**ement*, sendo "*pix*" a abreviação para "pictures". Representa o menor elemento de uma imagem digital exibida em um dispositivo (como por exemplo, a tela do monitor), ao qual pode-se atribuir uma cor (http://www.thefreedictionary.com/pixel).

um atributo do dado, como mostrado na Figura 2.12, que ao final, são arranjados conforme diferentes propósitos.

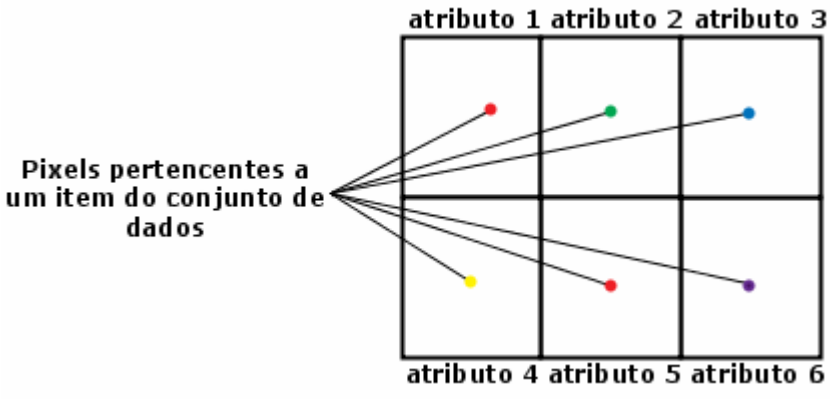

Figura 2.12: Exemplo da aplicação da técnica orientada a pixel para um dado com 6 atributos (adaptado de (Keim, 2000))

- Técnicas baseadas em dimensões – o espaço n-dimensional dos dados é dividido em subespaços que estão organizados e exibidos na forma hierárquica, caso os dados possuam esta característica, projetando esses espaços uns dentro dos outros, como ilustra a Figura 2.13. Técnicas baseadas no conceito de árvore podem ser utilizadas para representar dados com estruturas de hierarquia.

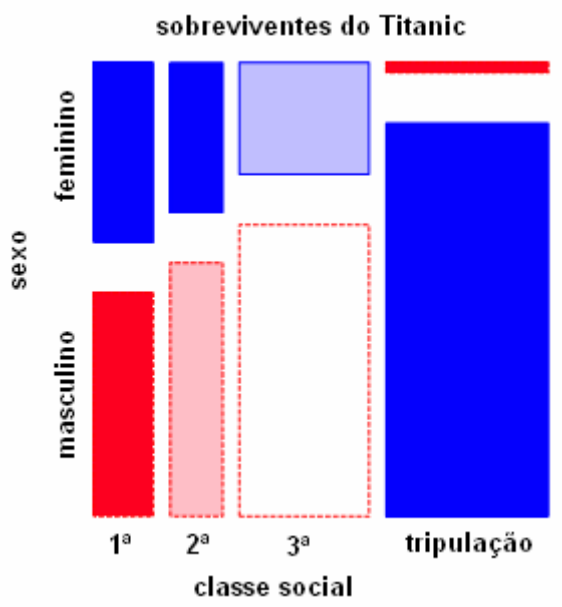

Figura 2.13: Exemplo de técnica que utiliza empilhamento de dimensões (adaptado de (Friendly, 1998))

3) Classificação pelas técnicas de interação e distorção

As técnicas de interação permitem ao usuário interagir diretamente com a visualização dos

dados de acordo com os objetivos da descoberta de conhecimento, enquanto as técnicas de distorção provêem meios para visualizar em maior ou menor nível os detalhes de uma porção de dados, sem perder a visão geral. Dentre essas técnicas para a exploração de dados, têm-se:

- Projeções dinâmicas – troca as projeções gráficas dinamicamente conforme a exploração do conjunto de dados.
- Filtragem interativa – faz o particionamento dos dados para focar somente os elementos de interesse.
- Zoom interativo – permite a visualização de um subconjunto de dados com vários níveis de detalhamento.
- Distorção interativa – consiste em mostrar um alto nível de detalhamento para uma parte dos dados, enquanto que para o restante dos dados exibidos, o nível de detalhamento é menor.
- Ligação e Seleção interativas – o objetivo é combinar diferentes métodos de visualização para cobrir as desvantagens que cada método possui, resultando em maiores informações que podem ser retiradas da visualização em comparação à utilização de somente uma técnica de visualização.

A Figura 2.14 mostra uma visão geral sobre as classificações discutidas anteriormente, sendo na Figura 2.14.a ilustrada uma comparação entre a classificação de Shneiderman (1996) e Keim (2002) quanto ao tipo de dados; na Figura 2.14.b, quanto ao tipo de tarefa, que Keim (2002) denomina por técnicas de interação e distorção; e na Figura 2.14.c, a classficação dada por Keim (2002) quanto ao tipo de técnica de visualização.

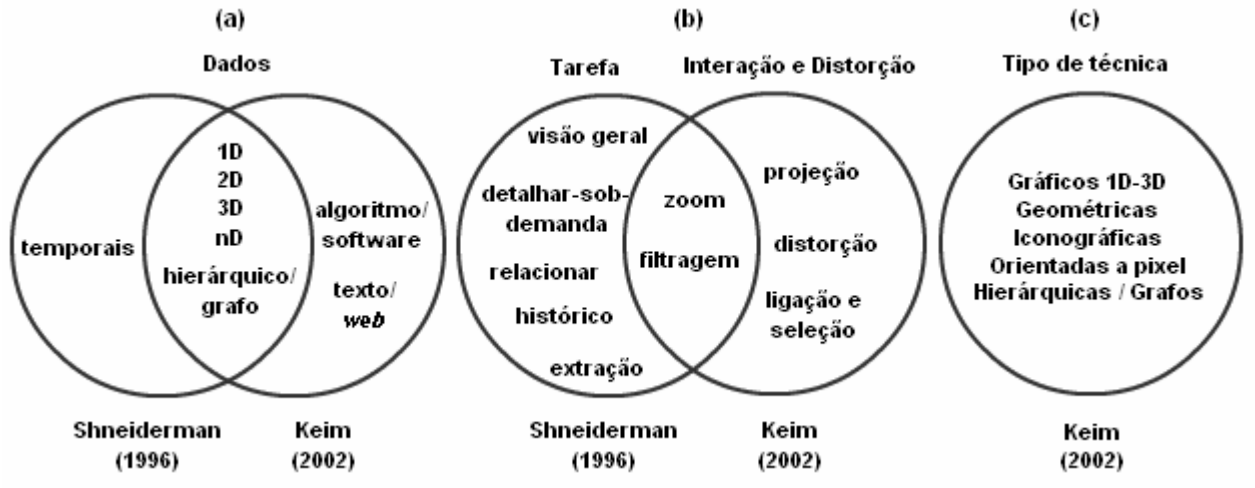

Figura 2.14: Comparação entre as taxonomias de técnicas de visualização de Shneiderman (1996) e Keim (2002)

c) Mineração Visual de Dados

Mineração Visual de Dados refere-se à aplicação de técnicas de visualização para avaliar, monitorar e guiar o processo de mineração de dados. Segundo Ganesh et al. (1996), essa avaliação consiste de exemplos de treinamento, exemplos de teste e modelos de aprendizado para verificar os resultados dos algoritmos de mineração. O monitoramento inclui atividades como rastrear o progresso dos algoritmos de mineração de dados, avaliando a relevância dos padrões no contexto das atualizações sofridas pelo banco de dados. A condução do processo de mineração de dados é influenciada pelas atividades realizadas pelo usuário que pode alterar os dados de entrada como também aprender novos padrões descobertos nos dados que somam conhecimentos ao tomador de decisão.

Segundo Kopanakis e Theodoulidis (2003) a mineração visual de dados diz respeito ao emprego de representações gráficas dos dados por todo o processo da mineração de dados, apresentada na Seção 2.2. O conceito de mineração visual de dados dado por estes autores é semelhante à classificação feita por Ankerst (2001) quanto ao nível de integração entre técnicas de visualização e mineração de dados, discutida na Seção 2.2.3.

Reforçando então a ideia de Ankerst (2001), Kopanakis e Theodoulidis (2003), a mineração visual pode ser aplicada sobre os dados brutos, a fim de realizar a manutenção quanto às discrepâncias e inconsistências presentes nos dados, bem como selecionar o conjunto de dados que seja relevante à extração de conhecimento. Dessa forma, a aplicação da visualização sobre dados brutos fornece maior clareza na formulação de hipóteses e auxilia na definição dos objetivos no processo de KDD.

A visualização também pode guiar a construção de um modelo de dados, definindo seus parâmetros e validando-o por meio de exemplos de dados para teste. Além disso, a visualização desse modelo permite a sua melhor compreensão pela equipe envolvida no processo de mineração.

Quando a visualização é empregada durante o estágio da validação, esta pode ser assumida por um gráfico de apresentação, uma vez que a informação já foi retirada dos dados e precisa apenas ser apresentada às partes interessadas. O papel da visualização neste caso é mapear o resultado obtido para um gráfico.

Mendonça e Sunderhaft (1999) afirmam ainda que a mineração visual propicia a combinação de atividades de visualização e mineração de dados em uma única ferramenta, permitindo a interação do analista com as visualizações utilizando as funcionalidades disponíveis na ferramenta como, por exemplo, *zoom*, rotações e controle nos níveis de detalhes dos dados.

Dessa forma, o analista tem condições de explorar os dados e encontrar padrões interessantes por meio da navegação sobre as imagens dos dados geradas pela ferramenta de mineração visual.

Assim, trabalhos relacionados à mineração visual de dados discorrem sobre uma das três categorias apresentadas anteriormente. Dentre esses, Hofmann et al. (2000) demonstram como o diagrama de mosaicos pode ser usado na visualização de regras de associação, que é uma tarefa pertencente à mineração de dados. Bruzzese e Buono (2004) apresentam a utilização de grafos e das coordenadas paralelas para também visualizar as regras de associação sob diferentes perspectivas. Hinneburg et al. (1999) combinam técnicas de visualização com algoritmo de agrupamento aplicado em base de dados multidimensionais. Ankerst et al. (1999) apresentam um método de classificação com árvores de decisão baseado em técnicas de visualização orientadas a pixel.

d) Visualização do Conhecimento

A definição do conhecimento geralmente é dada pela distinção entre os conceitos de dados, informação e conhecimento (Tuomi, 1999). Davenport e Prusak (1998) definem dados como um conjunto quantificável de elementos que representam um evento. No contexto de uma organização, podem ser expressos por registros de transações. Quando um significado é atribuído aos dados dentro de uma contextualização, obtém-se a informação. Da correlação entre informações e da interpretação sobre elas deriva-se o conhecimento. De acordo com Kock Jr. et al. (1996), a informação combinada com o conhecimento inserido em um domínio promovem uma ação, como ilustra a Figura 2.15. A ação pode ser, por exemplo, a interferência em um processo para que o desenvolvimento de um produto seja mais eficiente ou que tenha mais qualidade (Davenport e Prusak 1998).

O compartilhamento e disseminação do conhecimento é um fator importante para a gestão do conhecimento nas organizações (Silva, 2004). Segundo Eppler e Burkhard (2004), a área da Visualização da Informação possui várias técnicas desenvolvidas para a representação visual de dados, entretanto, não houve a preocupação de associar essas técnicas ao ato de disseminar o conhecimento entre os tomadores de decisão de uma organização. Assim, a Visualização do Conhecimento faz uso do potencial da visualização para melhorar a criação e o compartilhamento do conhecimento entre pelo menos duas pessoas.

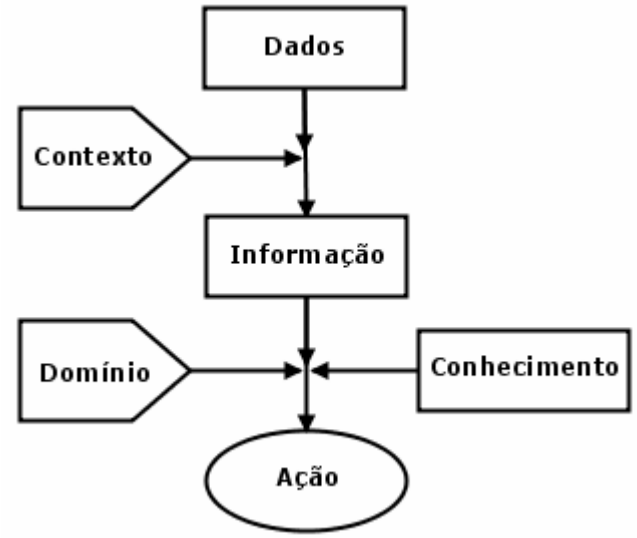

Figura 2.15: Relação entre dados, informação e conhecimento (adaptado de Kock Jr. et al., 1996)

Eppler e Burkhard (2004) propõem um framework para a visualização do conhecimento, ilustrada pela Figura 2.16, que segue quatro perspectivas a serem consideradas quando a visualização tem o propósito de disseminar conhecimento (Burkhard, 2005):

- Tipo de função que indica qual o objetivo da visualização;
- Tipo de conhecimento que precisa ser visualizado;
- A quem destina-se o conhecimento;
- Método de visualização utilizado.

Tipo de Função	Tipo de Conhecimento	Tipo de Receptor	Tipo de Visualização
Coordenação	O quê	Individual	Rascunho
Atenção	Como	Grupo	Diagrama
Lembrete	Por quê	Organização	Imagem
Motivação	Onde	Rede de comunicação	Mapa
Elaboração	Quem		Objeto
Nova compreensão			Visualização Interativa
			História

Figura 2.16: Framework de Visualização do Conhecimento (adaptado de Burkhard, 2005)

Segue a descrição destas quatro perspectivas para o framework de visualização do conhecimento, segundo (Burkhard, 2005).

O Tipo de Função é baseado no objetivo da percepção visual que distinguem-se em seis funções:

1) Coordenação: a representação visual é utilizada para coordenar a comunicação entre os indivíduos.

2) Atenção: prove avisos para manter a atenção, foca na identificação de padrões, tendências e transmissão de conhecimento.

3) Lembrete: melhora a capacidade de recordar e assim incentivar a aplicação de novos conhecimentos.

4) Motivação: estimula a interpretação e exploração do gráfico.

5) Elaboração: o processo de visualização do conhecimento promove a concepção de ideias e a interação com a visualização.

6) Nova compreensão: a visualização do conhecimento pode revelar conexões de informações antes ocultas, proporcionando novos entendimentos sobre o domínio.

O Tipo de Conhecimento é especificado em cinco classes, conforme as dimensões do conhecimento.

1) Conhecimento informativo: baseado em dados e informações – responde “o quê”;

2) Conhecimento de processo: conhecimento sobre determinado processo de um domínio – responde “como”;

3) Conhecimento experimental: conhecimento de causa-efeito, como evento de exceção – responde “por quê”;

4) Conhecimento de local: onde as informações podem ser encontradas – responde “onde”;

5) Conhecimento de pessoal: indivíduo que possui o conhecimento – responde “quem”.

O Tipo de Receptor procura identificar o contexto de quem recebe o conhecimento, que pode ser um indivíduo, uma equipe, uma organização (ou uma cultura), ou uma rede de comunicação entre diferentes organizações. Saber o tipo de receptor é importante na escolha do tipo de visualização apropriado para transferir conhecimento.

O Tipo de Visualização define o método de visualização por meio do qual será feita a representação do conhecimento. Utiliza qualquer tipo de representação visual como rascunhos, diagramas, imagens, objetos, visualizações interativas (baseadas em computador) e mapas como instrumentos de comunicação, que constituem a classificação dos tipos de visualização feita por Eppler e Burkhard descritas a seguir.

- Rascunhos (*Sketches*): São ideias desenhadas de maneira versátil para auxiliar no debate de um projeto ou pensamento. Por ser de fácil acesso, podem ser utilizados como meio de transmitir conhecimento entre as pessoas de uma organização, de modo mais flexível.
- Diagramas: Possuem representação esquemática que categorizam os dados, reduzindo a complexidade sobre eles, de modo que seja possível visualizar de forma mais clara a relação entre os seus valores, aumentando a compreensão sobre a informação.
- Imagens: Imagens podem provocar emoções e iniciar discussões a respeito de um tema; são expressivas e chamam a atenção; facilitam o aprendizado. Podem ser utilizadas para descrever metáforas visuais, que representam uma nova ideia a partir de elementos familiares ao receptor, como elementos da natureza, objetos, atividades ou conceitos abstratos, utilizados com um significado diferente para expressar um conceito. A árvore, por exemplo, é um conceito muito utilizado para explicar uma estrutura de dados hierárquicos.
- Mapas: Seguem as convenções da cartografia para a construção da visualização que refere-se a um conhecimento. Geralmente consiste de duas partes: uma camada que representa o contexto geral (uma rede ou cidade, por exemplo) e os elementos individuais mapeados nesse contexto. O mapa que descreve as linhas ferroviárias de uma estação de metrô é um exemplo dessa classe.
- Objetos: Objetos físicos ou modelos arquiteturais podem ser utilizados para explorar o conhecimento de uma composição espacial, explorando a visualização tridimensional, sobre o qual o conhecimento de um projeto pode ser concebido sob diferentes pontos de vista. Este tipo de visualização é adequado em exibições de arquiteturas ou modelos, como estátuas de dinossauro em um museu de ciências, ou maquetes de edifícios.
- Visualizações interativas: Permitem ao usuário explorar, manipular e interagir com visualizações de dados projetadas por um computador, pelas quais novos conhecimentos podem ser obtidos.
- História ou Imagens mentais: Constituem-se de visualizações criadas na mente dos indivíduos promovidas por histórias acerca do conhecimento compartilhado pela instituição.

e) Visualização de Dados

Chen et al. (2008) distinguem o emprego dos gráficos para duas finalidades: para apresentação de informações ou para a exploração de dados. Dessa forma, existem os gráficos de apresentação, utilizados para a apresentação e explicação de resultados, e os gráficos de exploração, que são construídos para auxiliar a exploração de dados. Segundo Chen et al. (2008), o gráfico de apresentação tem uma forma única de representação dos dados, por isso deve ser bem escolhido, para que passe a informação de maneira correta. O gráfico de exploração, por sua vez, tem por objetivo buscar resultados, encontrar informações e gerar ideias. Um gráfico para apresentação de informações é feito para ser visto por vários leitores, em contrapartida, vários gráficos de exploração são construídos para a investigação dos dados por um analista. A Visualização de Dados surge então como um novo campo de pesquisa no qual a visualização é aplicada como instrumento de inferência sobre os dados, assim as informações contidas nos dados devem ser reveladas por métodos adequados de visualização.

Quando trata-se de gráficos de exploração, Inselberg (2008) afirma que a transição dos dados para visualização deve seguir os seguintes critérios:

1) preservação da informação: deve ser possível a reconstrução de um conjunto de dados a partir de um gráfico, sendo cada valor de dado recuperável pela visualização, como por exemplo, no caso de gráficos de matriz de dispersão ou de coordenadas paralelas;
2) baixa complexidade representativa: o custo computacional de construir o gráfico deve ser baixo;
3) considerar qualquer dimensão: a dimensão dos dados não pode ser um fator limitante para a construção do gráfico;
4) tratar cada atributo uniformemente: contempla a arbitrariedade da distribuição dos atributos dos dados para as propriedades de uma representação gráfica. É uma característica típica de técnicas iconográficas, em que as distribuições dos atributos para as propriedades do ícone podem ser realizadas de modo arbitrário gerando, normalmente, resultados distintos;
5) exibir invariância sobre as projeções das transformações: o conjunto de dados original deve ser reconhecível após rotações, translações, escalas e mudanças de perspectiva. Matrizes de dispersão e coordenadas paralelas possuem esta característica;
6) revelar relações multivariadas do conjunto: segundo Inselberg (2008), o mais importante é poder reconhecer as relações entre os objetos por meio da visualização, em vez de apenas exibí-los.

7) ser baseado em uma metodologia matemática e algorítmica rigorosa: dessa forma elimina-se a ambiguidade nos resultados.

f) Processo Analítico Visual (*Visual Analytics*)

O Processo Analítico Visual tem como foco a extração de conhecimento contido em dados crus e heterogêneos, utilizando a visualização como intermédio no processo analítico, no qual existe a combinação da capacidade cognitiva humana e o poder computacional (Keim et al., 2008). Dessa forma, o Processo Analítico Visual está centrado nos interesses do usuário, que participa e orienta um processo fortemente interativo com métodos e técnicas para manipular os dados de modo que o conhecimento seja adquirido diretamente sobre eles.

Analogamente aos princípios definidos por Shneiderman (1996) para a Visualização da Informação, conhecidos como *Visual Information Seeking Mantra*, Keim et al. (2008, p. 171) realizaram uma adaptação para o Processo Analítico Visual: "*Analyze first, Show the Important, Zoom, filter and analyse further, Details on Demand*".

Tendo como base este princípio definido por Keim et al. (2008), pode-se sumarizar a solução proposta pelo Processo Analítico Visual, na qual a visualização é utilizada para representar os dados do modo mais fiel possível para, a partir de então, iniciar-se um ciclo de interação do usuário com os dados por meio de técnicas de: visualização, análise de dados (estatística e matemática), gerenciamento de dados, interação humano-computador, percepção e cognição.

Durante este processo, o usuário pode trabalhar com diferentes tipos de representações gráficas que facilitam o entendimento sobre os dados, paralelamente à avaliação de hipóteses levantandas em iterações anteriores (Keim et al., 2008). Assim, a visualização é usada como um elemento de interação, por meio da qual pode-se acompanhar a evolução do processo de extração do conhecimento, e reger as demais áreas relacionadas, que são:

- Análise de dados: ferramentas e métodos de extração de conhecimento amplamente pesquisados no campo de KDD e Mineração de Dados;
- Gerenciamento de dados: Uma pré-condição necessária para qualquer análise de dados é uma base de dados consistente e integrada. Segundo Keim et al. (2008), cabe ao sistema de gerenciamento de base de dados o papel de prover de modo eficaz os dados a serem analisados, uma vez que atualmente sua função está além do armazenamento, sendo responsável por manipular diversas classes de dados, como tipos numéricos, gráficos, textos, hipertextos, áudios e vídeos;

- Interação humano-computador: a interatividade pregada pelo Processo Analítico Visual necessita das técnicas desenvolvidas na área de estudo da interação humano-computador, dando suporte à integração entre as ações do usuário e da máquina. Uma interface apropriada beneficia a percepção e a cognição do usuário sobre os dados analisados;
- Percepção e cognição: técnicas advindas da psicologia, sociologia e neurociência contribuem para compor as técnicas de interação e distorção de um sistema para representação gráfica de dados, de modo que as funcionalidades combinem a comunicação gráfica e a interface com usuário de forma interativa;
- Infra-estrutura e avaliação: A construção de um sistema reunindo as disciplinas citadas requer o apoio de uma infra-estrutura que apoie na avaliação das soluções encontradas. Segundo Keim et al. (2008), isto deve ser feito a exemplo da área de Visualização da Informação, que possui pesquisas sobre taxonomias de técnicas considerando o tipo de dado e tarefa, consolidando um conhecimento a respeito de resultados esperados com a utilização de técnicas de visualização. Assim, a ideia é expandir esses parâmetros para o campo do Processo Analítico Visual. Scholtz (2006) traz uma abordagem sobre algumas métricas e metodologias para avaliar um ambiente construído para o processo analítico visual.

Desse modo, este processo busca uma solução integrada para suporte à tomada de decisão, combinando à visualização fatores humanos e técnicas de análise de dados. Segundo Keim et al. (2008), essa combinação visa identificar o melhor algoritmo para as tarefas de análise que podem ser automatizadas e, quando não houver esta possibilidade de automação, encontrar uma solução integrada envolvendo visualização e técnicas de interação. A Figura 2.17 ilustra a ideia do Processo Analítico Visual, que tem por objetivo (Keim et al., 2008):

- Sintetizar informação e extrair conhecimento de dados massivos, dinâmicos, ambíguos e muitas vezes conflitantes.
- Detectar o esperado e descobrir o inesperado.
- Prover avaliações eficazes para a tomada de decisão

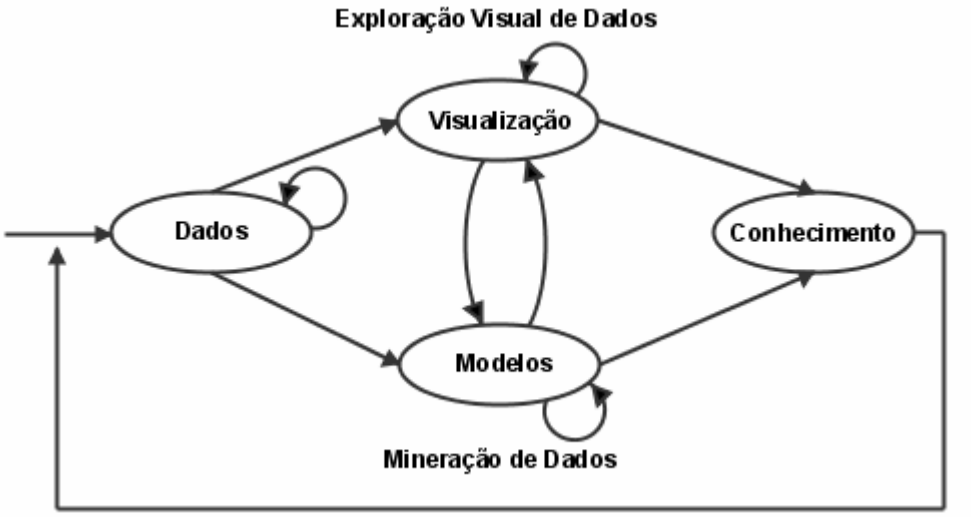

Figura 2.17: Forte integração entre os métodos visuais e automáticos de análise de dados para o suporte interativo à tomada de decisão (adaptado de Keim et al., 2008; p. 163)

O Processo Analítico Visual é um campo amplo e genérico, cujos princípios podem ser empregados nos mais variados domínios de aplicação. Keim et al. (2008) citam exemplos como aplicações para análise de rotas por meio do GPS (sistema de posicionamento global), no qual grande quantidade de dados é coletada por esse sistema. A função da aplicação apresentada é prover agrupamentos dos locais mais frequentados e o analista seleciona os locais de interesse. O sistema "interpreta" e indica uma rota num contexto espacial. Outro exemplo dado por estes autores, é uma aplicação para a análise da estruturas de redes aéreas, na qual o analista pode atribuir semântica aos níveis de visualização sobre o grafo formado pelo tráfego aéreo entre as cidades.

2.4. Trabalhos relacionados

O objetivo deste trabalho é fornecer embasamento para a escolha de técnicas de visualização que, se aplicadas diretamente sobre os dados a fim de obter algum tipo de conhecimento, torna-se opcional recorrer para demais métodos de mineração de dados. Para tanto, esta dissertação apóia-se em estudos desenvolvidos sobre análises ou avaliações de técnicas de visualização. Assim, esta seção é dedicada a apresentar, de forma breve, alguns trabalhos relacionados. Foram consideradas pesquisas como a de Oliveira e Levkowitz (2003), Herman et al. (2000) e Chen et al. (2008), apresentadas na Seção 2.4.1, por apresentarem uma visão geral das técnicas de visualização. As análises feitas por esses autores contribuem no sentido de identificar as situações em que determinada técnica é utilizada.

Embora o trabalho de Dias (2002) e de Rabelo (2007), apresentados na Seção 2.4.2, estejam no âmbito da mineração de dados, as avaliações realizadas por estes autores colaboram na elaboração desta dissertação no sentido de serem um guia para a captação dos parâmetros a serem analisados ao selecionar uma técnica de visualização.

Keim e Kriegel (1996) consideram o processo de mineração de dados como uma forte integração entre o especialista com a visualização completa dos dados, na qual estão aliados o poder cognitivo humano em formular hipóteses, selecionar e rejeitar padrões identificados por meio das visualizações manipuladas durante o processo de extração de conhecimento.

Nas subseções seguintes, os trabalhos citados acima são apresentados com maior abrangência, divididos em dois grupos: aqueles que tratam de análise e aqueles que tratam de avaliação de técnicas de visualização.

2.4.1. Análise de técnicas de visualização

Oliveira e Levkowitz (2003) apresentam um panorama geral sobre a Visualização da Informação no contexto do processo de KDD e mineração de dados, destacando trabalhos que abrangem o papel de técnicas de visualização na exploração e análise de grande quantidade de dados. Tais técnicas são abordadas de acordo com as taxonomias definidas por Shneiderman (1996), Keim e Kriegel (1996) – apresentadas na Seção 2.3.3 desta dissertação – e Card et al. (1999). Esses autores também descrevem softwares que implementam algumas das técnicas de visualização e outras propostas de ferramentas para a visualização de informação.

O trabalho desenvolvido por Herman et al. (2000) também constituiu-se de uma visão geral sobre técnicas de visualização, cujo tema principal são os grafos, sendo o emprego desse tipo de representação visual analisado sob diferentes perspectivas.

A demonstração do estado da arte da visualização de dados pode ser encontrada na obra de Chen et al. (2008) que reúne uma coleção de trabalhos de especialistas, organizados em capítulos que tratam dos princípios, metodologias e aplicações de técnicas de visualização de dados. Quando é apresentada a parte dos princípios, diversos autores discutem sobre determinadas técnicas de visualização como gráficos para modelos estatísticos, técnicas iconográficas, grafos e como algumas delas podem ser combinadas. Em seguida, na parte de metodologia, estão trabalhos relacionados às áreas de pesquisas como geografia (cartografia e mapas) e estatística, que utilizam a visualização para a análise de dados. E, por fim, os assuntos tratados na parte de aplicações abordam a visualização como principal meio de análise de dados em diferentes áreas como a medicina e aplicações comerciais e industriais.

2.4.2. Avaliação de técnicas de visualização

Nos trabalhos de Keim e Kriegel (1996) e Keim (2000), é apresentada a técnica de visualização orientada a pixel. Para avaliar a capacidade desta técnica, foi realizada uma comparação entre técnicas de visualização pertencentes às classes geométrica, iconográfica, orientadas a pixel, hierárquica e grafos. Essa comparação foi realizada segundo os seguintes critérios:

- características dos dados: número de atributos, quantidade de dados, adequação a valores categóricos;
- tipo de tarefa: capacidade de identificar agrupamentos;
- características da visualização: grau de sobreposição dos dados e curva de aprendizagem.

Outro trabalho envolvendo avaliação de técnicas que podem ser utilizadas no processo KDD é o de Dias (2002). Nele, a autora estabelece parâmetros para a escolha de técnicas de mineração de dados. Embora não esteja diretamente relacionado com técnicas de visualização, alguns critérios adotados por Dias para a definição desses parâmetros são suscetíveis de serem considerados também na escolha de técnicas de visualização de dados. Aspectos como o tipo de dados, desempenho computacional, funcionalidades pertinentes à técnica escolhida, são fatores a serem ponderados na escolha de uma técnica de visualização.

Já o trabalho de Rabelo et al. (2008) constitui-se em uma avaliação de técnicas de visualização de dados para fornecer subsídios aos analistas de sistemas KDD na escolha de uma técnica de visualização adequada. Foram avaliadas técnicas geométricas e iconográficas utilizadas para representar visualmente os resultados de algoritmo de agrupamento durante a mineração de dados. Além disso, foram aplicadas técnicas de visualização sobre os dados brutos, com o objetivo de identificar inconsistências existentes na base de dados e também detectar padrões ou tendências nos dados originais. Rabelo et al. (2008) consideraram os seguintes parâmetros para a avaliação das técnicas de visualização geométricas e iconográficas:

- Escalabilidade – suporte a quantidade de dados
- Dimensionalidade – suporte a quantidade de atributos
- Representação de tipos de dados que distinguem-se em:
  - qualitativos nominais ou ordinais;
  - quantitativos discretos ou contínuos;
  - híbrido (qualitativo e quantitativo).
- Interatividade – possibilidade de inserção de técnicas de interação
- Interpretabilidade – facilidade de extrair informação.

- Relacionamento entre os atributos.
- Correlação – grau de relacionamento entre os atributos.

2.5. Considerações Finais

A Visualização é uma área multidisciplinar, pois pode ser empregada concomitantemente com outras técnicas de extração de conhecimento ou, ainda, a própria visualização pode prover informações suficientes para obter o conhecimento necessário. Devido à eficácia que a visualização possui na representação de dados, várias áreas de aplicações utilizam técnicas de visualização para representar ou transmitir conhecimento.

Na Análise Exploratória de Dados, os gráficos são utilizados para representar os resultados de uma avaliação estatística; na Visualização da Informação, técnicas mais elaboradas foram desenvolvidas para auxiliar o processo de descoberta de conhecimento, aprimorando o entendimento sobre os resultados parciais ou finais da mineração; a Mineração Visual de Dados, por sua vez, alia técnicas de visualização e de mineração de dados convergindo para uma ferramenta visual analítica; por outro lado, a Visualização do Conhecimento amplia as alternativas de representação gráfica, buscando retratar e transmitir o conhecimento de um indivíduo ou grupo; a Visualização de Dados trata a visualização não apenas como uma ferramenta auxiliar no processo de extração de conhecimento, mas sim como o principal recurso dentro desse processo; por fim, o Processo Analítico Visual generaliza a aplicabilidade da visualização, empregada para analisar e gerenciar dados, ao mesmo tempo em que facilita a compreensão do usuário sobre os dados por meio da interatividade oferecida pela ferramenta de visualização.

Diante de todas as versões de terminologias apresentadas sobre a visualização, há um ponto em comum entre todas elas: o papel do especialista é de fundamental importância, pois é ele quem decidirá se a visualização lhe acrescenta algum conhecimento útil ou não. Portanto, é desejável que a escolha de uma técnica de visualização seja feita criteriosamente, de modo a tirar melhor proveito do que os dados têm a dizer.

3. Teoria Fundamentada em Dados

3.1. Considerações iniciais

Neste capítulo é apresentada a Teoria Fundamentada em Dados (do inglês *Grounded Theory*), que é um método de pesquisa utilizado no presente trabalho. A razão de destinar um capítulo a essa metodologia é devido ao pouco conhecimento a seu respeito e rara utilização na área da computação ou informática. Além disso, suas etapas são diferentes dos métodos científicos clássicos, que normalmente constituem-se de formulação do problema, definição da hipótese, verificação da hipótese e conclusão.

A teoria fundamentada em dados (TFD), por sua vez, não parte de nenhuma hipótese, buscando estruturar uma teoria que sustenta-se sobre uma análise sistemática dos dados coletados durante a pesquisa, os quais podem ser baseados em referências bibliográficas e/ou conclusões a respeito de questionários e/ou entrevistas respondidas por um grupo de pessoas ou, ainda, observações sobre o ambiente que constitui o foco do trabalho.

De modo geral, as etapas desta metodologia são: formação da amostragem teórica, codificação dos dados, diagramação e formulação da teoria (Rodrigues et al., 2004; Coleman e O'Connor, 2007; Matavire e Brown, 2008; Rodon e Pastor, 2007).

Segundo Simoni e Baranauskas (2003) e Wainer (2007) uma pesquisa qualitativa pode seguir, além da TFD, outras metodologias, tais como: estudo de caso e pesquisa-ação.

De acordo com Simoni e Baranauskas (2003), o estudo de caso é um dos métodos mais utilizados na área da computação e de sistemas de informação, por focar no estudo do processo e nas opiniões dos participantes dentro de um contexto real e particular. As principais fases de um estudo de caso, segundo a definição de Kitchenham et al. (1995), são: definição de hipótese; seleção de um projeto piloto; seleção de métodos de comparação; minimização dos fatores de confusão; planejamento do estudo de caso; monitoramento do estudo de caso; análise e apresentação dos resultados.

No estudo de caso, a condução de toda a pesquisa baseia-se na definição da hipótese. Um projeto piloto é utilizado para representar o objeto de estudo do trabalho, que pode ser, por exemplo, a proposta de um novo processo ou modelo. Os resultados obtidos da análise do estudo de caso devem ser comparados com outros resultados de métodos diferentes. Um estudo de caso pode estar sujeito a fatores de confusão, que estão relacionados às condições dos partipantes (curva de aprendizado, motivação) como também à especificidade do domínio em questão. Tais fatores devem ser minimizados ou contornados para manter a confiabilidade dos resultados. O planejamento do estudo de caso define os critérios, operações e medições a serem realizadas durante o experimento e, no monitoramento, deve-se assegurar que as atividades planejadas serão cumpridas corretamente. Por fim, a análise dos resultados pode ser apresentada por meio de métodos estatísticos.

A pesquisa-ação é um método de pesquisa originário da Psicologia, que busca modificar o ambiente sob estudo, pondo em prática propostas de soluções aos problemas (Wainer, 2007). O pesquisador possui um papel ativo e altamente interativo com o processo e com os participantes envolvidos. As principais etapas da pesquisa-ação, segundo Filippo (2008), são: planejamento; ação sobre o problema; observação e reflexão sobre os resultados obtidos. Estas etapas são apresentadas em formato de um ciclo, que podem ser repetidas diversas vezes, fazendo com que o conhecimento advindo de ciclos anteriores seja refinado nos seguintes.

De acordo com Filippo (2008), na fase de planejamento define-se a solução que será investigada, quais ações serão tomadas, quais dados a serem coletados e como a análise sobre esses dados será feita. A fase da ação corresponde à aplicação da solução planejada. Na fase de observação, são coletados os dados definidos na etapa de planejamento, que podem referir-se ao momento anterior e/ou posterior à aplicação da solução. A fase de reflexão consiste da análise dos efeitos decorrentes da aplicação da solução bem como na reflexão sobre quais refinamentos precisam ser feitos.

A Tabela 3.1, baseada principalmente nos trabalhos de Rodon e Pastor (2007), Simoni e Baranauskas (2003) e Filippo (2008), resume as principais características dos métodos de pesquisa qualitativa mencionados e suas respectivas etapas.

Tabela 3.1: Comparação entre os métodos de pesquisa qualitativa citados neste trabalho

Metodologia	**Características**	**Etapas**
Teoria Fundamentada em Dados	Baseia-se em um processo indutivo para construir uma teoria por meio de uma ação contínua entre coleta e análise dos dados da qual emerge a teoria	Formação da amostragem teórica Coleta dos dados Codificação Diagramação Formulação da teoria
Estudo de caso	Consiste do estudo de um cenário individualizado, observado pelo pesquisador. Neste cenário busca-se avaliar um método ou ferramenta dentro de um contexto real, considerando inclusive a opinião dos sujeitos envolvidos.	Definição de hipótese Seleção de um projeto piloto Seleção de métodos de comparação Minimização dos fatores de confusão Planejamento do estudo de caso Monitoramento do estudo de caso Análise e apresentação dos resultados
Pesquisa-ação	Baseia-se na introdução de mudanças dentro de uma organização ou comunidade e a observação dos efeitos causados.	Planejamento Ação sobre o problema Observação Reflexão sobre os resultados

Este trabalho encontra-se no campo da pesquisa qualitativa e como tal, necessita de uma metodologia de avaliação qualitativa para neutralizar a subjetividade nos resultados da pesquisa. O objetivo principal deste trabalho, que é a definição de diretrizes para escolher técnicas de visualização, não enquadra-se no método Estudo de Caso em razão de não haver uma hipótese prévia para o tipo de problema tratado, uma vez que as diretrizes emergem de um estudo investigativo sobre as técnicas de visualização. Também não enquadra-se nos princípios da Pesquisa-ação, que visa modificar diretamente o ambiente estudado.

Como não houve uma definição prévia de uma hipótese acerca dos parâmetros identificados e tão pouco das diretrizes, apesar de feita a revisão bibliográfica, a teoria envolvida na definição das diretrizes foi, de fato, construída conforme foram sendo analisados artigos e materiais coletados de periódicos e eventos promovidos por instiuições e academias. Dessa forma, a TFD é o método de pesquisa qualitativa adequado à proposta desta dissertação. Assim, em seguida são descritas as principais características deste método.

3.2. Princípios da Teoria Fundamentada em Dados

Este método foi proposto em 1967 pelos sociólogos Glaser e Strauss, como alternativa ao método científico tradicional, que é baseado em definição de hipótese, técnicas de verificação e análises quantitativas (Rodrigues et al., 2004; Coleman e O'Connor, 2007 e Matavire e Brown, 2008). Constitui-se na explicação de fatos observados em dados coletados e analisados sistematicamente, resultando em uma teoria próxima à "realidade", por ser derivada a partir de

hipóteses que emergem do estudo e análise dos dados coletados durante a pesquisa, originados da literatura ou de entrevistas em campo. Dessa forma, de acordo com Haig (1995), a teoria emerge dos dados por meio de indução.

Os dados coletados passam por um processo de codificação, no qual são questionados em relação a o que, quando, como, quais as consequências, etc., procedimento denominado comparação constante (Dantas et al., 2009). A partir destas simples perguntas é possível elaborar os primeiros conceitos e, em uma etapa posterior, organizar esses conceitos em categorias. Da correlação entre essas categorias é possível moldar uma teoria que vai sendo refinada durante esse processo.

Assim, à medida que novos dados são coletados, as informações levantadas até então devem ser re-analisadas para garantir a consistência das categorias que emergiram durante esse processo. Quando novos dados não agregam informações relevantes para a construção dos conceitos e das categorias tem-se, nesse ponto, a saturação teórica e o fechamento da amostragem teórica.

Dessa forma, de acordo com Rodrigues et al. (2004), Coleman e O'Connor (2007) e Matavire e Brown (2008), os principais elementos que constituem a Teoria Fundamentada em Dados são: amostragem teórica, codificação dos dados e memorandos. A codificação dos dados ainda pode ser distinta entre codificação aberta, codificação axial e codificação seletiva. Rodon e Pastor (2007) ainda citam mais um elemento, denominado diagramação. Cada componente da TFD é descrito a seguir.

- Amostragem teórica – refere-se à definição do assunto da pesquisa e ao processo de coleta de dados. Durante a formação da amostragem teórica, a coleta e a análise de dados são realizadas concomitantemente. O investigador constantemente compara os dados uns com os outros e os agrupa em categorias, conforme os conceitos atribuídos aos dados analisados.
- Codificação aberta – nomeação de conceitos por meio de palavras ou expressões que formam códigos preliminares para a categorização dos dados. Nessa fase os dados são comparados em termos de semelhanças e diferenças.
- Codificação axial – os dados são reorganizados relacionando categorias e subcategorias definidas na codificação aberta.
- Codificação seletiva – nesta etapa é eleita uma categoria central pela qual as demais são refinadas e a teoria é detalhada em torno desta.

- Memorandos – são registros em forma de anotações criados durante o processo de coleta e análise de dados, que são utilizados como apoio à formulação da teoria. As anotações transcrevem as ideias do pesquisador, feitas por meio de declarações, hipóteses ou questionamentos (Coleman e O'Connor, 2007).
- Diagramação – consiste da ilustração dos conceitos e categorias emergentes durante a etapa de codificação como apoio à explicação da teoria emergente. Segundo Rodon e Pastor (2007), não há uma forma sistemática para construir esse diagrama.

3.2.1. Diferentes vertentes da Teoria Fundamentada em Dados

Originalmente, a teoria fundamentada em dados foi proposta por Glaser e Strauss em obra entitulada *The Discovery of Grounded Theory*. Posteriormente as visões desses autores divergiram e surgiram duas vertentes para a metodologia. Walker e Myrick (2006) e Matavire e Brown (2008) fazem uma profunda discussão sobre as particularidades dessas duas abordagens. De acordo com Walker e Myrick (2006, p. 550), ambas as versões utilizam: codificação, comparação constante, questionamentos, amostragem teórica e memorandos.

Basicamente, a diferença entre as concepções de Glaser e Strauss está na posição epistemológica (Rodrigues et al., 2004; Walker e Myrick, 2006; Coleman e O'Connor, 2007). Segundo esses autores, a versão de Glaser enfatiza que a teoria deve surgir da indução e criatividade do pesquisador perante a análise dos dados; a versão de Strauss, formulada juntamente com Corbin, defende que a teoria deve surgir de uma análise sistemática dos dados por meio de procedimentos bem definidos (Walker e Myrick, 2006). Além disso, na versão da metodologia atribuída a Glaser, qualquer consulta à literatura deve ser evitada para que o pesquisador não seja influenciado por nenhuma pré-concepção a respeito do assunto estudado. Na linha defendida por Strauss e Corbin, por outro lado, a literatura deve ser consultada conforme a necessidade do pesquisador (Rodon e Pastor, 2007).

3.2.2. Pontos Críticos da Teoria Fundamentada em Dados

A TFD é um método indutivo, em que não há uma hipótese pré-definida a qual se quer confirmar ou refutar. Ao contrário, a hipótese ou teoria é gradualmente estruturada ao longo da investigação sobre os dados coletados, os quais podem ser obtidos por meio de entrevistas, formulários e literatura, sendo esta fase denominada coleta de dados.

Segundo Rodrigues et al. (2004), a TFD deve ser aplicada de forma objetiva, de modo que outro pesquisador possa confirmar os resultados sobre os mesmos dados, demonstrando que eles falam por si só e estão alheios à subjetividade do analista.

Entretanto, existe a questão do ponto de saturação teórica que, segundo Rodrigues et al. (2004), não está claramente definido pelos autores da teoria. Isto simboliza uma dificuldade para o pesquisador, que deve sentir confiança e convicção suficiente para cessar a coleta de dados.

Além disso, ainda segundo Rodrigues et al. (2004), a validade do método pode ser comprometida dependendo do rigor aplicado na metodologia. Assim, desde que o analista seja rigoroso nos seus métodos de análise, seguindo os critérios estabelecidos pela vertente escolhida, as conclusões alcançadas são significativas e válidas, pois emerge de uma realidade empírica evidenciada pelos próprios dados.

3.3. Teoria Fundamentada em Dados aplicada à área da Informática

Wainer (2007, p.28) discute a pesquisa qualitativa utilizando exemplos de trabalhos oriundos da área de Sistemas de Informação que, apesar de não ser uma subárea da Computação, está relacionada a campos tradicionais da mesma como, por exemplo, engenharia de software, interfaces humano-computador e sistemas colaborativos. Segundo Wainer (2007), Sistemas de Informação tem como foco solucionar problemas relacionados à gestão de informação, que envolve problemas práticos quanto ao uso e desenvolvimento de sistemas de informação, análise sobre os impactos econômicos, financeiros e sociais desses sistemas nas organizações e, também, os efeitos na adoção ou no desenvolvimento de novos sistemas. Por esses motivos, Wainer (2007) afirma que a área de Sistemas de Informação está mais próxima a departamentos de Administração e Engenharia de Produção.

De fato, trabalhos como os de Orlikowski (1993), Esteves et al. (2002) e de Oliveira et al. (2009) são exemplos da aplicação da TFD para resolver problemas ligados a Sistemas de Informação. Em Orlikowski (1993), esta metodologia é utilizada para demonstrar a adoção de ferramentas CASE pelas organizações ao longo do tempo. Esteves et al. (2002) discutem a aplicação da TFD analisando quatro trabalhos relacionados a Sistemas de Informação que utilizaram esta metodologia. Oliveira et al. (2009) baseiam-se na TFD para explicar o procedimento de alocação de pessoas em projetos de software.

Matavire e Brown (2008) fazem um levantamento da produção acadêmica da área de Informática que empregam a Teoria Fundamentada em Dados em seus trabalhos publicados em

periódicos como *Information Technology & People*, *Journal of Management Information Systems*, *Information Research* e *MIS Quarterly*.

Analisando esses trabalhos, Matavire e Brown (2008) concluem que existem quatro abordagens adotadas pelos autores: (1) a vertente defendida por Glaser; (2) a versão de Strauss e Corbin; (3) a aplicação dos métodos de análise da TFD e; (4) uma metodologia híbrida de Glaser e Strauss. A abordagem predominantemente adotada, de acordo com Matavire e Brown (2008), é a aplicação dos métodos de análise para a validação da pesquisa.

3.4. Teoria Fundamentada em Dados como método de avaliação adequada a esta dissertação

A Teoria Fundamentada em Dados é adotada neste trabalho por permitir ao pesquisador desenvolver uma explicação teórica acerca de um assunto fundamentando-se em avaliações empíricas e observações sobre os dados (Orlikowski, 1993).

Neste trabalho, as avaliações empíricas são realizadas por meio da utilização de ferramentas computacionais de visualização sobre exemplos de base de dados como modo de confirmar os parâmetros levantados durante pesquisa na literatura, constituindo a triangulação. São observadas as características do conjunto de dados analisados de acordo com os parâmetros identificados para, posteriormente, escolher uma técnica de visualização mais adequada.

A coleta de dados é baseada principalmente na literatura, como sugere Dick (2005), sendo considerada neste trabalho como fonte de dados. A análise comparativa é realizada da mesma forma daquela em que dados são originários de entrevistas ou formulários. Os artigos e materiais analisados durante este trabalho são em sua maioria sobre experiências e/ou experimentos sobre as técnicas de visualização.

Segundo Wainer (2007), a novidade na área de Ciência da Computação é fundamental, tanto que, dependendo da área ou subárea, a apresentação de um novo algoritmo, modelo, programa ou sistema e sua comparação com alternativas existentes são fatores suficientes para a sua aceitação em periódicos da área de Informática.

O trabalho de Inselberg e Dimsdale (1990), por exemplo, está relacionado a uma nova técnica de visualização para a época, no caso, as coordenadas paralelas. Um modelo matemático foi descrito como método de validação do trabalho.

O trabalho de Keim e Kriegel (1996) é outro exemplo de trabalho que apresenta as técnicas orientadas a pixel como uma nova alternativa frente às técnicas de visualização

existentes. Estes autores realizaram uma comparação com as técnicas figuras de arestas e coordenadas paralelas para demonstrar as vantagens das técnicas orientadas a pixel.

Quando Chernoff, em 1973, apresentou uma técnica de representação visual baseda em características do rosto humano, denominada Faces de Chernoff, este autor o fez por meio de experimentos empíricos (Chernoff, 1973).

Assim, a partir de trabalhos como os citados anteriormente, outros foram desenvolvidos, como trabalhos de: Shimabukuro (2004), que utilizou as técnicas orientadas a pixel; Fanea et al. (2005), que desenvolveram uma técnica que combina coordenadas paralelas e *star glyphs*; e Klippel et al. (2009), que estudaram a influência das características de *star glyphs* na tarefa de classificação.

Todo esse conjunto de publicações forneceu informações necessárias para a identificação dos parâmetros que influenciam na utilização de técnicas de visualização e, consequentemente, a definição das diretrizes para escolher técnicas de visualização adequadas às características dos dados. O papel da Teoria Fundamentada em Dados neste trabalho foi de prover os métodos para que estes objetivos fossem alcançados.

Um preceito importante da Teoria Fundamentada em Dados é que, para gerar resultados precisos e úteis, é necessário que o contexto do objeto em estudo seja racionalizado em categorias formadas por meio de agrupamentos de conceitos, de forma que expliquem determinado fenômeno (Rodrigues et al., 2004; Hunter et al., 2005). Allan (2003) discute dois métodos de codificação: micro-análise e pontos-chave.

A codificação por micro-análise consiste da fragmentação do texto palavra por palavra ou linha por linha e atribui-se um significado a cada segmento de frase, como exemplifica a Tabela 3.2, que ilustra a aplicação da micro-análise sobre o texto de uma entrevista. Esse método apresenta-se vagaroso, por ter que percorrer todas as linhas do texto para encontrar informações relevantes e, também, inconsistente, pois uma frase pode ser fragmentada de diversas formas e cada formato pode ser relacionado a significados diferentes. Além disso, o segmento de texto isolado pode ser interpretado de maneira diferente do que quando visto no texto original.

Tabela 3.2: Exemplo de codificação utilizando micro-análise - adaptado de (Allan, 2003)

Texto da entrevista	**Código**
do meu ponto de vista	Perspectiva pessoal
a principal mudança é	Declaração
na mudança de tecnologia	Mudança de tecnologia
ou a melhoria do produto	Mudança no produto
realizado pelo fornecedor	Declaração Mudança no fornecedor
não pode garantir que	Declaração Incerteza

Na codificação por pontos-chave, em vez de codificar trechos do texto, são analisados os conceitos que podem ser abstraídos de um fragmento de texto. A Tabela 3.3 mostra um exemplo de utilização desse método de codificação. Dessa forma, segundo Allan (2003), os conceitos já surgem dos dados, que posteriormente podem ser reorganizados em categorias, que uma vez relacionadas e organizadas constroem a teoria. No caso do presente trabalho, os conceitos e as categorias identificadas geram as diretrizes indicativas para técnicas de visualização conforme as características apresentadas pelo conjunto de dados analisados.

Tabela 3.3: Exemplo de codificação utilizando pontos-chave – adaptado de (Allan, 2003)

Pontos-chave	**Código**
O Manual de Normas de gerenciamento de configuração foi criado a partir do um estudo de requisitos para melhorar o controle de desenvolvimento de software.	Controle de software Software desenvolvimento
A empresa X exige um sistema de gerenciamento de configuração com o menor custo possível	Requerimento para um sistema de gerenciamento de configuração
A sincronização de mudanças foi considerada como uma parte essencial e inerente ao processo de software	Mudanças Processo de software

Pelas abordagens sobre a TFD apresentadas nos trabalhos de Rodrigues et al. (2004), Walker e Myrick (2006), Coleman e O'Connor (2007) e Rodon e Pastor (2007), pode-se dizer que este trabalho segue naturalmente a vertente de Strauss e Corbin, pelo fato da literatura ser, além de fonte de dados, o padrão para a comparação dos resultados obtidos com a utilização de ferramentas de visualização. Além disso, a abordagem de Glaser, discutida também pelos autores citados anteriormente, não condiz com os objetivos estabelecidos, por desaprovar uma consulta inicial à literatura.

Rodon e Pastor (2007) também seguem a abordagem de Strauss e Corbin, cujo processo seguido por estes autores, baseado na TFD, está ilustrado na Figura 3.1.

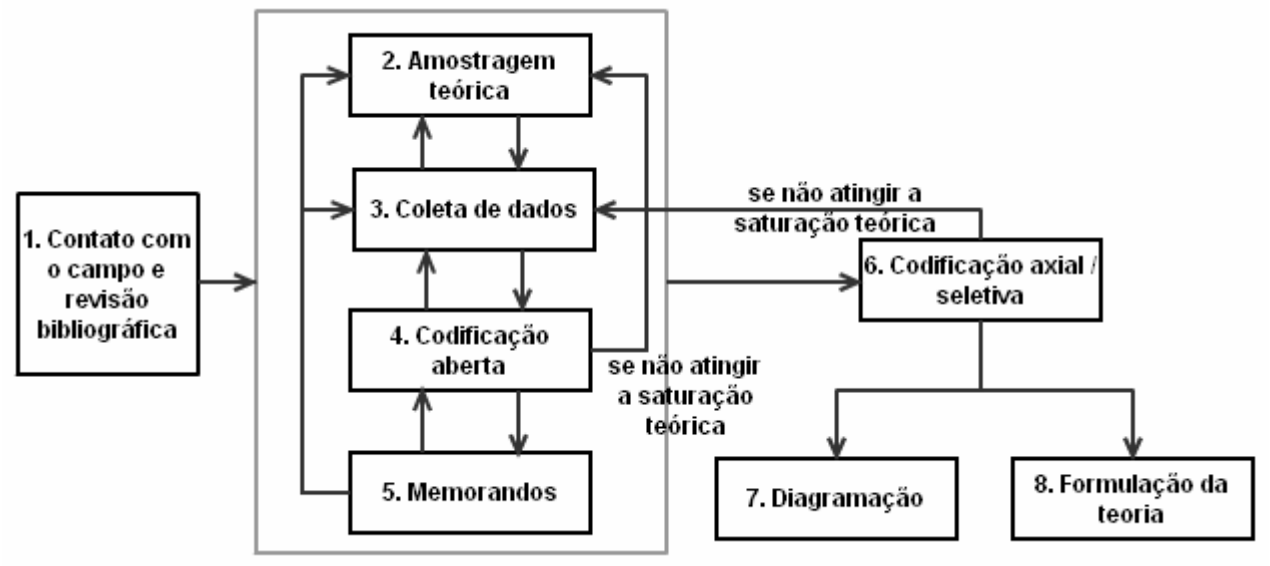

Figura 3.1: Versão da Teoria Fundamentada em Dados seguida por Rodon e Pastor (2007) (adaptado de (Rodon e Pastor, 2007))

A versão da TFD adotada neste trabalho segue um processo semelhante ao descrito na Figura 3.1, e também assemelha-se à versão da TFD discutida no trabalho de Orlikowski (1993). A Figura 3.2 apresenta, então, o processo metodológico seguido neste trabalho, baseado nessa metodologia.

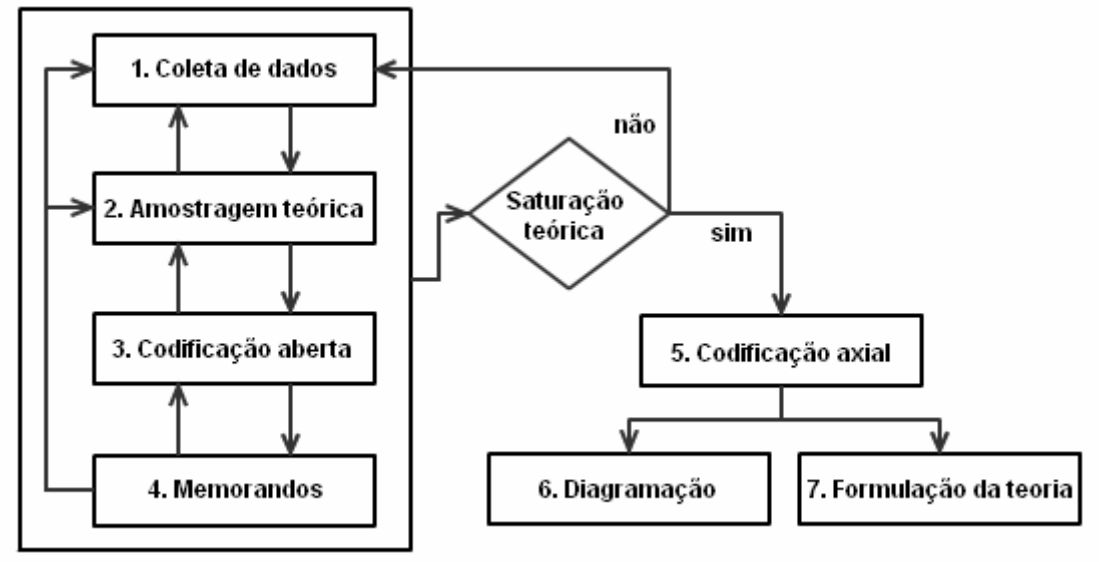

Figura 3.2: Processo metodológico adotado, baseado na Teoria Fundamentada em Dados

As diferenças entre a versão de Rodon e Pastor (2007) e a versão descrita na Figura 3.2 estão na formação da amostragem teórica e no processo de codificação. Rodon e Pastor (2007) constroem a amostragem teórica por meio da coleta de dados inicialmente sobre a revisão bibliográfica e, posteriormente, por meio de entrevistas a gerentes de organizações selecionadas. No caso deste trabalho, a amostragem teórica é contruída basicamente sobre a literatura existente, como Dick (2005) afirma que pode ser feito.

A segunda diferença reside no processo de codificação, no qual Rodon e Pastor (2007) consideram a codificação seletiva, enquanto que neste trabalho este tipo de codificação não é realizada, assim como no trabalho de Orlikowski (1993). A razão disto é que não há uma única categoria central que pode ser nomeada como o principal conceito para a formulação da teoria final. Neste trabalho, há um conjunto de fatores (parâmetros e tipos de técnicas) que geram as diretrizes indicativas para técnicas de visualização, conforme as características apresentadas pelo conjunto de dados analisados.

Portanto, neste trabalho, a TFD é utilizada como metodologia para a identificação dos parâmetros e composição das diretrizes. O procedimento de identificação dos parâmetros seguiu o método de pontos-chave, nos quais os conceitos, correspondentes aos parâmetros, foram investigados em trabalhos publicados em periódicos cujo assunto envolve técnicas de visualização, sendo este procedimento referente à etapa da codificação aberta. O passo seguinte consistiu da organização dos conceitos obtidos, realizada por meio de uma análise da relação entre esses parâmetros e as técnicas de visualização, perfazendo a codificação axial. Os memorandos foram retirados de observações feitas pelos autores dos trabalhos estudados, que ajudaram a compor as diretrizes para a escolha de técnicas de visualização, correspondendo à formulação da teoria.

3.5. Considerações Finais

A Teoria Fundamentada em Dados revelou-se uma metodologia adequada para este trabalho de dissertação, por ser uma metodologia indutiva, baseada na observação dos dados que são constantemente comparados e organizados em conceitos e categorias, para então ser possível a conclusão de uma teoria (ou fenômeno).

Por meio da apresentação desta metodologia neste capítulo, buscou-se esclarecer o encaminhamento desta pesquisa, que utiliza os métodos de análise da TFD para validar as diretrizes definidas no final deste trabalho.

Dessa forma, no Capítulo 4 é descrito como os conceitos identificados na fase de codificação aberta constituem os parâmetros envolvidos na escolha de técnicas de visualização e como esses parâmetros estão associados às categorias de técnicas de visualização, o que corresponde à fase de codificação axial.

No Capítulo 5 são apresentados exemplos de utilização de técnicas de visualização analisando os parâmetros definidos no Capítulo 4, constituindo o procedimento de triangulação, que reforça os conceitos levantados nas codificações aberta e axial, sendo possível, a partir desse ponto, a formulação das diretrizes descritas no Capítulo 6.

4. Parâmetros para a escolha de técnicas de visualização

4.1. Considerações iniciais

Neste capítulo são descritos os parâmetros identificados como sendo fundamentais na escolha de técnicas de visualização. A definição desses parâmetros foi realizada seguindo a metodologia da Teoria Fundamentada em Dados, discutida no Capítulo 3.

Inicialmente, conforme a TFD, o processo de codificação foi realizado seguindo o método de pontos-chave, conforme explicado na Seção 3.4, que possibilitou a identificação dos parâmetros que influenciam na escolha de técnicas de visualização na fase de codificação aberta. Na sequência, na fase de codificação axial, os parâmetros foram associados às técnicas de visualização.

Os pontos-chave foram retirados de textos das referências bibliográficas estudadas, sendo agrupados de acordo com os seguintes parâmetros nomeados: tipo de dado, tipo de tarefa, volume e dimensionalidade dos dados, e posição dos atributos. Estes parâmetros, por sua vez, foram analisados conforme as classes de técnicas de visualização, também identificadas na literatura como visto na Seção 2.3.3, sendo a taxonomia de Keim (2002) adotada neste trabalho, que divide as técnicas de visualização em: gráficos convencionais (para dados 1D a 3D); técnicas iconográficas; geométricas; orientadas a pixel; e técnicas baseadas em dimensões. A respeito desta última classe, é adotada a nomenclatura “técnicas hierárquicas ou baseadas em grafos”, que é a mesma terminologia utilizada em (Keim e Kriegel, 1996).

Assim, nas seções a seguir é mostrada uma análise das técnicas de visualização com o objetivo de identificar parâmetros para a sua escolha, são descritos os parâmetros identificados e, na sequência, é apresentada uma análise desses parâmetros em relação às técnicas de visualização.

4.2. Análise de Técnicas de Visualização

A Tabela 4.1 traz o sumário das técnicas discutidas neste trabalho. Foram escolhidas as técnicas de visualização encontradas com mais frequência na literatura estudada, sobre as quais foram destacadas as características principais de cada uma conforme os parâmetros levantados.

Tabela 4.1: Técnicas de visualização abordadas neste trabalho

Categorias		**Nome da técnica**
1D a 3D		Histograma
		Diagrama de caixa
		Gráfico de dispersão (*Scatter plot*)
		Gráfico de contorno
Multidimensionais	Iconográficas	Faces de Chernoff
		Star glyphs
		Figuras de aresta (*Stick figure*)
	Geométricas	Matriz de dispersão
		Coordenadas Paralelas
	Orientadas a pixel	*Query-dependent techniques*
		Query-independent techniques
Hierárquicas ou Baseadas em grafos		Grafos
		Cone trees
		Treemap
		Empilhamento de dimensões (*Dimensional stacking*)
		Gráfico de mosaico (*Mosaic plot*)

Ao todo, são analisadas cinco categorias de técnicas de visualização (gráficos 1D a 3D, iconográficas, geométricas, orientadas a pixel e hierárquicas ou baseadas em grafos) a partir de cinco parâmetros identificados (tipo de dado, tipo de tarefa, volume, dimensionalidade e posição dos atributos no gráfico). As Tabelas 4.2, 4.3, 4.4, 4.5, e 4.6 fornecem uma visão geral da análise realizada para as técnicas selecionadas de cada classe, descritas na Tabela 4.1, em relação aos parâmetros que devem ser considerados ao escolher representações visuais para um conjunto de dados disponíveis.

Estas tabelas sumarizam a análise dos parâmetros (denominados nesta seção como conceitos) em relação a cada categoria de técnica de visualização, realizando a codificação axial, conforme estabelece a Teoria Fundamentada em Dados. Na sequência, são detalhados o processo de obtenção desses parâmetros e sua análise em conjunto com as técnicas de visualização mencionadas por referências da literatura.

A Tabela 4.2 mostra as principais características das técnicas pertencentes à classe de gráficos 1D a 3D. As tarefas comumente realizadas pela utilização de técnicas dessa classe são relacionadas a análises estatísticas sobre a distribuição de frequência de um atributo ou a correlação entre determinados atributos. Geralmente os dados a serem visualizados são de natureza quantitativa e passam por processos de tratamento, tais como métodos de transformação e redução de dados (Myatt, 2007), acarretando, consequentemente, em um volume de dados pequeno. Para essa classe de técnica, o posicionamento dos atributos em geral não influencia na interpretação do gráfico, devido à dimensionalidade ser baixa e ao fato da locação dos atributos estar bem definida no gráfico, sendo o comportamento de um atributo analisado em função de um segundo e/ou terceiro atributo.

Tabela 4.2: Análise dos parâmetros em técnicas da classe de gráficos 1D a 3D

Categoria	**Conceitos**				
1D a 3D	**Tarefa**	**Tipo de dado**	**Dimensão**	**Volume**	**Posição**
Histograma	Análise da distribuição de frequência, verificação de padrões e detecção de *outliers*	Quaisquer dados qualitativos e quantitativos	1-D	Pequeno	Não influencia
Diagrama de caixa	Idem Histograma	Quantitativos contínuos	1-D		
Gráfico de dispersão	Verificar correlação entre dois atributos e detecção de *outliers*	Quantitativos contínuos	2-D		
Gráfico de contorno	Verificar correlação de um atributo em função de outros dois	Quantitativos discretos e contínuos	3-D		

Na Tabela 4.3 estão sumarizadas as principais características das técnicas iconográficas analisadas. Os atributos dos dados são mapeados para as propriedades de um ícone que representam cada registro do conjunto. A identificação de ícones semelhantes permite a formação de agrupamentos, e a detecção de elementos com formatos discrepantes das demais representações pode significar casos de *outlier*.

Ward (2008) apresenta uma discussão sobre o processo de mapeamento de atributos dos dados para os atributos visuais de um ícone. Para efeitos práticos, o autor considerou em seu trabalho somente valores numéricos escalares, sendo que atributos categóricos (ou qualitativos) e outros não-numéricos foram convertidos para uma forma numérica. Keim (2000) também avalia que técnicas iconográficas não são apropriadas para representar dados qualitativos. Além desses autores, Rabelo (2007) também admitiu dificuldade na representação desse tipo de dados ao utilizar técnicas iconográficas, sendo necessário realizar um processo de transformação dos dados para valores numéricos.

Quanto à representação do volume de dados, as Faces de Chernoff e *Star Glyphs* comportam um volume pequeno de dados (Oliveira e Levkowitz, 2003), pois cada ícone corresponde a um registro, o que limita a representação de um grande volume (Rabelo, 2007). Neste caso, modos de interação com o gráfico, tais como zoom, *brushing* (seleção de determinada região), podem ser utilizados para obter uma visualização mais apurada de um conjunto ou um único registro. Já a técnica de Figuras de Arestas, segundo Keim (2005), permite a visualização de um volume de dados maior em relação às Faces de Chernoff e *Star Glyphs*, podendo exibir uma quantidade de dados na ordem de 10^3. Portanto, o uso de técnicas iconográficas é recomendado para análise de dados quantitativos com volume pequeno a médio, com mais de três dimensões, em que técnicas convencionais de representação de dados 1D a 3D não suprem completamente a exploração e descoberta de padrões sobre esses dados.

Tabela 4.3: Análise dos parâmetros em técnicas da classe iconográfica

Categoria	**Conceitos**				
Iconográficas	**Tarefa**	**Tipo de dado**	**Dimensão**	**Volume**	**Posição**
Faces de Chernoff	Detecção de agrupamentos e *outliers*, identificação de padrões de comportamento (Bruckner, 1978)	Quantitativos discretos e contínuos	Até 18 atributos (Chernoff, 1973; Bruckner, 1978)	Pequeno	Não influencia
Star Glyphs	Detecção de agrupamentos e *outliers*	Quantitativos discretos e contínuos	Até 80 atributos (Rabelo, 2007)	Pequeno	Não influencia
Figuras de Aresta	Detecção de agrupamentos nos dados conforme a composição de diferentes texturas formadas pelos ícones.	Quantitativos discretos e contínuos	De 5 a 15 atributos (Keim, 2005)	Médio	Influencia

Das técnicas iconográficas apresentadas na Tabela 4.3, Figuras de Arestas é um exemplo em que a posição dos atributos pode influenciar na análise dos dados, quando realizada a tarefa de agrupamento, devido às famílias de ícones que podem ser derivadas a partir da variação da distribuição dos atributos dos dados para as propriedades do ícone, formando variadas representações (Pickett e Grinstein, 1988; p. 516).

As faces de Chernoff já possuem uma estrutura fixa para o ícone, cujas propriedades correspondem às características de um rosto humano e, desse modo, a variação das posições dos atributos não é um aspecto relevante para esta técnica. Mas há estudos sobre quais propriedades do ícone podem ser mais representativas para a interpretação dos resultados como, por exemplo, os olhos e o formato da face que são os aspectos que mais chamam a atenção (Morris et al., 2000; Lee et al., 2003).

Do mesmo modo, a técnica *Star Glyph* possui uma estrutura fixa para todas as estrelas que representam os registros, sendo o posicionamento dos atributos um fator que não influencia na análise dos dados, embora existam estudos (Peng et al., 2004; Klippel et al., 2009) sobre a melhor ordem de distribuição dos atributos para obter formatos de estrelas que facilitem a tarefa de agrupamento e classificação.

A Tabela 4.4 resume as principais características das técnicas geométricas mencionadas na Tabela 4.1. De forma geral, essas técnicas servem principalmente para a análise de correlação entre os atributos e a observação sobre o comportamento de um grupo de registros, que normalmente devem ser do tipo quantitativo. Coordenadas Paralelas também suportam dados do tipo qualitativo.

Um número de dimensões extremamente grande compromete a visualização para as técnicas geométricas apresentadas na Tabela 4.4. No caso da matriz de dispersão, a limitação quanto à dimensionalidade é devido ao espaço de exibição da tela para comportar a combinação dos atributos (Rabelo, 2007). Quanto às coordenadas paralelas, a representação de muitos atributos faz com que os eixos verticais fiquem muito próximos, dificultando a detecção de padrões. Inselberg (2008) afirma, no entanto, que nos casos em que há mais de centenas de atributos a serem visualizados, os indivíduos envolvidos no processo geralmente não possuem muita familiaridade com os dados, e acabam por incluir atributos supérfluos para a visualização. Nestas situações, é ideal selecionar somente aqueles de real interesse.

As técnicas geométricas podem ser aplicadas sobre um volume médio de dados. Inselberg (2008), por exemplo, afirma que conseguiu trabalhar com uma base com cerca de 9000 registros. Quanto maior o volume de dados, maior é a possibilidade de sobreposição dos elementos que representam cada registro, dificultando a detecção de agrupamentos e padrões. No caso das coordenadas paralelas, uma solução dada por Theus (2008) é a aplicação de saturação (grau de transparência das linhas poligonais de um determinado grupo de registros) ou a utilização de cores para o destaque de diferentes agrupamentos, que são aplicáveis também para a matriz de dispersão (Rabelo, 2007).

O posicionamento dos atributos é um fator que influencia na interpretação da visualização fornecida pelas Coordenadas Paralelas (Inselberg, 2008), mas não é relevante para a Matriz de Dispersão, uma vez que esta é composta por um conjunto de gráficos de dispersão, em que o posicionamento dos atributos está bem definido, sendo a tarefa verificar o comportamento de um atributo em relação aos demais.

Tabela 4.4: Análise dos parâmetros em técnicas da classe geométrica

Categoria	**Conceitos**				
Geométricas	**Tarefa**	**Tipo de dado**	**Dimensão**	**Volume**	**Posição**
Matriz de dispersão	Correlação entre atributos e detecção de agrupamentos por meio de cores	Quantitativos contínuos	Dezenas de atributos (em torno de 15, segundo Rabelo(2007))	Médio	Não influencia
Coordenadas Paralelas	Visão geral dos dados, identificação de agrupamentos, *outliers* e correlação entre atributos	Quantitativos e qualitativos	Centenas de atributos (Inselberg, 2008)	Médio	Influencia

A Tabela 4.5 sumariza os tipos de visualização abordados na categoria de técnicas orientadas a pixel. As tarefas principais são detecção de padrões e identificação de relacionamento nos dados. Técnicas dessa classe são próprias para visualizar dados do tipo quantitativo, com média a alta dimensionalidade e destacam-se por representar grande quantidade de dados. A sobreposição dos elementos representados no gráfico não influencia na análise da visualização, sendo este um fator diferencial das técnicas geométricas (Keim e Kriegel, 1996). O posicionamento dos pixels nas janelas depende do algoritmo aplicado para gerar a visualização.

Tabela 4.5: Análise dos parâmetros em técnicas da classe orientadas a pixel

Categoria	**Conceitos**				
Orientadas a pixel	**Tarefa**	**Tipo de dado**	**Dimensão**	**Volume**	**Posição**
Dirigido a dados *(Query-independent technique)*	Detectar padrões e correlação entre os atributos	Dados quantitativos (ex. dados temporais)	10 a 100 atributos (Keim, 2005)	Grande	Influencia
Dirigido a resultados *(Query-dependent technique)*	Verificar relacionamento dos atributos diante de um resultado de uma consulta (query)	Dados quantitativos (cujos atributos estão no contexto da consulta)			

A Tabela 4.6 sumariza as técnicas hierárquicas descritas na Tabela 4.1. Em geral, as técnicas hierárquicas ou baseadas em grafos servem para visualizar a estrutura de relacionamento entre os elementos, seja uma hierarquia ou uma rede. Além disso, existem técnicas dessa classe que podem representar dados qualitativos, como por exemplo, o Gráfico de Mosaicos, que possui algumas características da técnica *Treemaps*, mas com menos recursos, tais como áreas proporcionais ao valor do atributo, construção baseada em uma classificação inerente aos dados, e formato dependente da ordem dos atributos (Hofmann, 2008). Segundo Oliveira e Levkowitz (2003), a técnica de Empilhamento de Dimensões pode ser aplicada a dados armazenados em

tabelas, sem estarem necessariamente arranjados em um formato hierárquico, como assumem as técnicas *Cone trees* e *Treemaps.*

Técnicas hierárquicas podem representar dados com altas dimensões, considerando que os atributos podem ser representados pelos nós de um grafo. Quanto à quantidade de dados, de acordo com Oliveira e Levkowitz (2003), essas técnicas são adequadas para um volume que varia entre pequeno e médio, como pode ser constatado, por exemplo, nos trabalhos de Shneiderman (2006), Baehrecke et al. (2004) e Bederson et al. (2001). Estes últimos autores afirmam, inclusive, que um passo ambicioso no avanço da técnica *Treemaps* seria a capacidade de acomodar milhões de nós. A posição dos atributos nesse tipo de representação gráfica é um fator que geralmente deve-se prestar atenção, principalmente quando há níveis de hierarquia entre os elementos.

Tabela 4.6: Análise dos parâmetros em técnicas da classe hierárquica ou baseadas em grafos

Categoria	**Conceitos**				
Grafos	**Tarefa**	**Tipo de dado**	**Dimensão**	**Volume**	**Posição**
Grafos	Visão geral da estrutura de relação entre os elementos	Dados com estrutura relacional (hierarquia ou rede)	Alta	Pequeno a médio	Influencia
Cone Trees / Cam Trees	Facilitar a navegação pela estrutura de árvore	Adequado para a visualização de estruturas de arquivos e diretórios	Alta (pode exibir cerca de 1.000 dados)	Pequeno a médio	Não influencia
Treemap	Visualizar agrupamento de dados	Dados que apresentam alguma relação de hierarquia e/ou taxonomia	Alta (pode representar 2 a 20 agrupamentos)	Médio	Influencia
Gráfico de mosaico	Visão geral de atributos qualitativos	Qualitativos nominais e ordinais	Média	Pequeno	Influencia
Empilhamento de dimensões	Detecção de padrões, agrupamentos e *outliers*	Quantitativos discretos e contínuos	Média	Médio	Influencia

4.3. Parâmetros Identificados

Nesta seção são descritos os parâmetros a serem considerados na escolha de técnicas de visualização de dados, de acordo com a literatura pesquisada. Tais parâmetros são: tipo de dado, tipo de tarefa, volume e dimensionalidade dos dados, e posição dos atributos no gráfico da visualização.

A Tabela 4.7 ilustra a codificação aberta, procedimento que deu origem aos parâmetros mencionados, seguindo a estratégia de codificação por pontos-chave, descrita em (Allan, 2003). Na primeira coluna estão relacionados os conceitos abstraídos da literatura, cujas respectivas referências estão indicadas na coluna seguinte. Na terceira, estão os códigos definidos de acordo com os conceitos descritos, que correspondem aos parâmetros identificados, discutidos na sequência.

Tabela 4.7: Codificação por pontos-chave baseada na literatura

Pontos-chave	**Referência**	**Código**
As técnicas de visualização podem ser classificadas, dentre outros critérios, pelo tipo de dado a ser visualizado.	Shneiderman (1996) Freitas et al. (2001) Keim (2002)	Tipo de dado
Um dos aspectos considerados na classificação de técnicas de visualização é o tipo de tarefa, que fornece um meio de interação entre o analista e a visualização.	Shneiderman (1996) Keim (2002) Pillat et al. (2005) Valiati (2008)	Tipo de tarefa
As técnicas de visualização estão sujeitas a algumas limitações, como é o caso da quantidade de dados que determinada técnica pode exibir.	Keim e Kriegel (1996) Oliveira e Levkowitz (2003) Rabelo et al. (2008)	Volume
As técnicas de visualização também podem ser classificadas de acordo com o número de dimensões que podem ser visualizadas.	Shneiderman (1996) Grinstein et al. (2001) Keim (2002) Oliveira e Levkowitz (2003)	Dimensionalidade
Em algumas classes de técnicas, a disposição dos atributos no gráfico pode influenciar na interpretação da visualização, como é o caso da análise de correlação entre atributos, em que a distância relativa entre as variáveis é relevante para a observação.	Keim (2000) Morris et al. (2000) Ankerst (2001) Oliveira e Levkowitz (2003) Inselberg (2008) Klippel et al. (2009)	Posição dos atributos

4.3.1. Tipo de dado

Um fator que deve ser levado em consideração para a escolha de uma técnica de visualização é o tipo de dado manipulado. Segundo a taxonomia de Shneiderman (1996), os dados podem ser classificados em unidimensionais (1D), bidimensionais (2D), tridimensionais (3D), temporais, multidimensionais, hierárquicos ou com estruturas de grafos. Keim (2002) acrescenta dados texto/web e algoritmos/software como classe de dados. Além disso, os dados ainda podem ser classificados em:

- Dados quantitativos – referentes a valores numéricos, que distinguem-se em duas categorias:
  - Contínuos – abrangem um intervalo numérico representado por infinitos valores associados a uma determinada escala e mensurados por algum instrumento. Exemplos dessa classe estão nas unidades de medida como peso, altura, velocidade, tempo, etc.
  - Discretos – compreendem um número que representa o valor de uma contagem, como por exemplo, quantidade de livros de uma biblioteca, número de peixes em um aquário, número de filhos, entre outros.
- Dados qualitativos (ou categóricos) – representam uma classificação que atribui um nome ou rótulo de acordo com as características do atributo, que dividem-se em:
  - Ordinais – apesar de não serem numéricos, obedecem a uma relação de ordem, como o grau de escolaridade, avaliação atribuindo conceitos como ótimo, bom, regular ou ruim, temperatura (frio, morno, quente), etc.
  - Nominais – não existe uma ordenação entre as categorias, dessa forma, só é possível definir uma relação de igualdade ou diferença quanto à classe, como por exemplo, sexo, estado civil, cor dos olhos, entre outros.

A Tabela 4.8 a seguir, baseada em Keim (2002), Freitas et al. (2001) e Shneiderman (1996), traz o panorama geral dos tipos de dados considerados na literatura. Desse modo, segundo Freitas et al. (2001), os dados podem ser classificados de acordo com critérios referentes à classe (ou tipo) de informação, natureza e dimensão (número de atributos). A dimensionalidade dos dados já foi discutida na Seção 2.3.3, sob as perspectivas de Keim (2002) e Shneiderman (1996). A natureza do domínio refere-se às propriedades primitivas dos atributos, distintas entre tipos quantitativos e qualitativos, como mencionadas acima. Porém, nem todos os dados enquadram-se especificamente numa classificação de acordo com a dimensionalidade ou natureza. Esses podem ser descritos em termos de sua classe de informação, como os dados do tipo texto/web, que representam o texto ou conteúdo multimídia em uma página *web*, que não são passíveis de serem visualizados por técnicas mais simples. Os dados também podem corresponder a informações sobre o relacionamento entre seus registros, comumente representados por grafos ou estruturas hierárquicas. Além disso, algoritmos ou códigos-fonte de programas são exemplos de dados que podem ser visualizados por técnicas visuais específicas que elevam o entendimento do desenvolvedor sobre o software (Keim, 2002).

Tabela 4.8: Sumário da caracterização dos dados

Critério	**Classe**	**Exemplo**
Classe de informação	Texto/ web	Documentos, textos (que podem estar na *web*)
	Hierárquico e grafos	Chamadas telefônicas e páginas da Internet
	Algoritmos e softwares	Dados de depuração do código, como *logs*
Natureza do domínio	Qualitativos nominais	Gênero
	Qualitativos ordinais	Estágio de um conhecimento (inicial, intermediário, avançado)
	Quantitativos discretos	Total de vendas de um produto
	Quantitativos contínuos	Pressão arterial
Dimensão	1D	Dados com característica de tempo (dados temporais)
	2D	Dados geográficos, mapas, superfície de terreno
	3D	Dados do mundo real: dados médicos, dados de construção de edifícios
	nD	Tabelas de banco de dados

4.3.2. Tipo de tarefa

Um dos critérios de classificação de técnicas de visualização considerado por Shneiderman (1996) e Keim (2002) é o tipo de tarefa. Shneiderman (1996) identifica sete tarefas que podem ser realizadas pelos usuários sobre a visualização, como visto na Seção 2.3.3: (1) Visão geral; (2) Zoom; (3) Filtragem; (4) Detalhar sob demanda; (5) Relacionar dados; (6) Histórico das ações realizadas; (7) Extração de subconjuntos de dados e parâmetros de consulta. Keim (2002), por sua vez, define uma classificação de técnicas de visualização, apresentada na Seção 2.3.3, considerando funcionalidades de técnicas de interação e distorção utilizadas para interagir com a representação gráfica, que podem ser: Projeção, Filtragem, Zoom, Distorção e Ligação & Seleção (*Link* & *Brush*).

Pillat et al. (2005) e Valiati (2008) apresentam outras taxonomias de tarefas que o usuário ou analista podem realizar de acordo com objetivos em relação à utilização de uma técnica de visualização. Diante disso, neste trabalho são consideradas as tarefas mais comuns, tais como:

- Visão geral dos dados: visualizar os dados completamente;
- Verificação da correlação entre os atributos: o grau de relacionamento entre as variáveis pode revelar padrões de comportamentos e tendências;
- Identificação de regras, padrões e características importantes;
- Identificação de agrupamentos: atributos que possuem comportamentos semelhantes;
- Detecção de *outliers*: conjunto de dados que apresentam valores discrepantes ou atípicos em relação ao restante dos dados.

4.3.3. Volume e dimensionalidade dos dados

Diante do fato de existirem grandes quantidades de dados armazenados nas organizações e considerando a possibilidade do uso de técnicas de visualização na busca por padrões nesses dados, a capacidade que uma determinada técnica suporta em relação ao volume e dimensionalidade (quantidade de atributos) dos dados deve ser um dos fatores preponderantes na sua escolha.

Segundo Rabelo (2007), não há um consenso de uma quantia de atributos que determina se os dados possuem alta ou baixa dimensão. Oliveira e Levkowitz (2003) afirmam que a interpretação de dados com mais de cinco dimensões torna-se difícil para o ser humano. Desse modo, estes autores sugerem, para fins gerais, que dados com dimensionalidade baixa envolvam até quatro atributos; média, de cinco a nove; e alta, acima de dez, como resume a Tabela 4.9. Estes autores discutem ainda a definição de volume dos dados, que geralmente carregam expressões como “pequena” ou “grande quantidade de dados”. Da mesma forma, para fins gerais, são adotados neste trabalho os termos “pequeno, “médio” e “grande” para designar o volume de dados, distintos de acordo com as ordens de grandeza relacionadas na Tabela 4.10.

Tabela 4.9: Unidades de grandeza consideradas para designar a dimensionalidade dos dados

Dimensionalidade	**Número de atributos**
Baixa	≤ 4
Média	5 a 9
Alta	≥ 10

Tabela 4.10: Unidades de grandeza consideradas para designar o volume dos dados

Volume	**Ordem de grandeza**
Pequeno	10^1 a 10^2 = 10 a 100
Médio	10^3 a 10^5 = 1.000 a 10.000
Grande	10^6 a 10^7 = 1.000.000 a 10.000.000

4.3.4. Posição dos atributos

Este parâmetro está efetivamente mais relacionado às técnicas de visualização do que às características dos dados. Durante a análise visual, dependendo da técnica de visualização utilizada, alterar os lugares ocupados pelos atributos no gráfico pode ser uma ação significativa para o processo de descoberta de conhecimento.

Para a tarefa de análise de correlações, por exemplo, a disposição na qual os atributos são visualizados é relevante, uma vez que esta correlação torna-se mais perceptível quando os

atributos de interesse estejam em posições próximas um do outro.

Na técnica Coordenadas Paralelas (Inselberg, 2008), por exemplo, os atributos são mapeados para eixos verticais paralelos ligados por meio de linhas poligonais que representam um registro do conjunto de dados. Conforme esses eixos são trocados de lugar, as linhas poligonais são desenhadas de diferentes formas, o que pode revelar padrões ou correlações entre os atributos. Por esse motivo, Inselberg (2008) recomenda a interação do analista com os eixos das coordenadas paralelas durante a exploração dos dados.

Na tarefa de agrupamento, diferentes ordens de mapeamento dos atributos para as propriedades do gráfico podem gerar resultados que contribuam para identificar grupos de dados com características ou comportamentos semelhantes. É o caso da técnica *Star Glyph*, sobre a qual Klippel et al. (2009) discutem a ordem de atribuição dos atributos para as propriedades dos ícones. Morris et al. (2000), apresentam um estudo sobre a eficácia das faces de Chernoff para detecção de agrupamentos, indicando quais elementos da face podem contribuir melhor para este tipo de tarefa.

Ankerst et al. (1998) propuseram um algoritmo heurístico para avaliar a similaridade entre os atributos e então posicioná-los próximos um do outro para facilitar a visualização do relacionamento entre eles. Este algoritmo foi aplicado para técnicas orientadas a pixel e também para coordenadas paralelas.

A Figura 4.1 ilustra um exemplo de aplicação do algoritmo proposto por Ankerst et al. (1998), na qual à esquerda (Figura 4.1.a), estão representados os registros pela técnica orientada a pixel denominada *circle segments*, e à direita (Figura 4.1.b) os pixels foram rearranjados conforme uma tarefa de classificação e, consequentemente, atributos similares estão próximos no gráfico, facilitando a visualização de grupos.

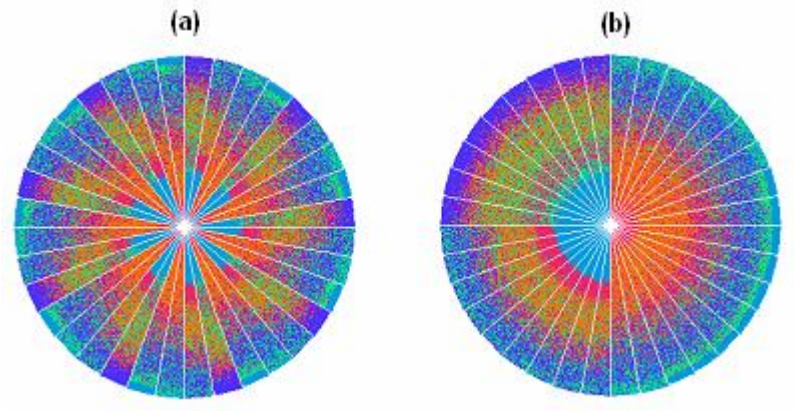

Figura 4.1: Exemplo de classificação dos atributos pela similaridade de comportamento. (a) Visualização dos registros por meio da técnica circle segment; (b) Rearranjo dos pixels após aplicação da técnica de classificação (adaptado de Ankerst, 2001)

4.4. Análise dos Parâmetros Identificados em Relação às Técnicas de Visualização

Após terem sido identificados os parâmetros que podem influenciar na escolha de técnicas de visualização, nesta seção é demonstrada a associação desses parâmetros às técnicas de visualização por meio de memorandos retirados da literatura. Isto perfaz a codificação axial, da TFD, procedimento no qual os conceitos sobre os parâmetros para a escolha de técnicas de visualização foram associados e organizados juntamente às classes de técnicas de visualização. São apresentados exemplos de técnicas na ordem da taxonomia definida por Keim (2001), que classifica as técnicas de visualização como: gráficos convencionais (para dados de 1D a 3D), iconográficas, geométricas, orientadas a pixel e hierárquicas ou baseadas em grafos, conforme apresentado na Seção 2.3.3. Cada técnica de visualização é analisada de acordo com a seguinte ordem dos parâmetros: (I) tipo de tarefa; (II) tipo de dado; (III) dimensionalidade; (IV) volume e; (V) posicionamento dos atributos.

4.4.1. Gráficos 1D a 3D

Dados com um a três atributos podem ser visualizados por gráficos comumente utilizados pela Estatística, na Análise Exploratória de Dados, conforme discutido na Seção 2.3.3. A seguir, são descritos alguns gráficos como exemplos de representação de dados 1D a 3D, como (a) histograma; (b) gráficos de dispersão; (c) gráfico de contorno; (d) diagrama de caixa; e (e) outras técnicas pertencentes a essa classe. Há uma grande variedade de gráficos estatísticos, como pode ser verificado em (NIST / SEMATECH, 2006).

a) Histograma

O histograma tem como propósito principal representar graficamente a distribuição de frequência de um atributo (NIST/SEMATECH, 2006), sendo composto por barras verticais justapostas que são posicionadas sobre um plano cartesiano. Cada barra corresponde a uma classe cuja base representa o intervalo de valores pertencente a uma escala e a altura corresponde à frequência.

a.I) Tipo de tarefa

Por meio do histograma, é possível verificar: ponto central dos dados (média, mediana); tipo de distribuição (normal, simétrica, assimétrica); presença de *outliers*; presença de moda (classe(s) com maior frequência).

a.II) Tipo de dado

Pode representar dados quantitativos (discretos e contínuos) e qualitativos (ordinais e nominais) (Myatt, 2007; p.52). A seguir, são descritos alguns tipos de histogramas que podem revelar características presentes nos dados. (Figuras retiradas do material disponível em <www.dcce.ibilce.unesp.br/~adriana/ceq/.../histograma.doc>, acesso em 30 de novembro de 2009).

- Simétrico – a frequência é mais alta no centro e decresce gradualmente para as caudas de modo simétrico, tendo o formato de um sino, como ilustra a Figura 4.2, significando que a média aproxima-se da mediana. Caracteriza processos padronizados e estáveis.

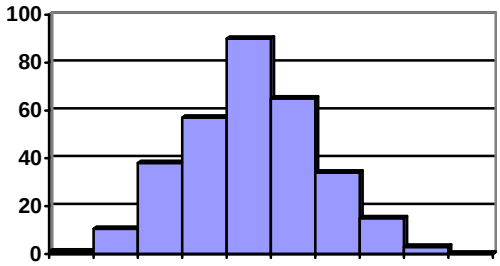

Figura 4.2: Exemplo de histograma simétrico

- Assimétrico (com apenas um pico) – a frequência varia bruscamente, formando pico à esquerda ou à direita, produzindo uma cauda mais longa, como pode ser visto no exemplo ilustrado pela Figura 4.3. Se a inclinação estiver à esquerda, significa que a média é menor que a mediana; se a inclinação estiver à direita, a média é maior que a mediana. Pode caracterizar processos que, por exemplo, apresentam algum tipo de anormalidade que ao decorrer do tempo tende a decrescer ou crescer, correspondendo aos picos na esquerda ou na direita, respectivamente.

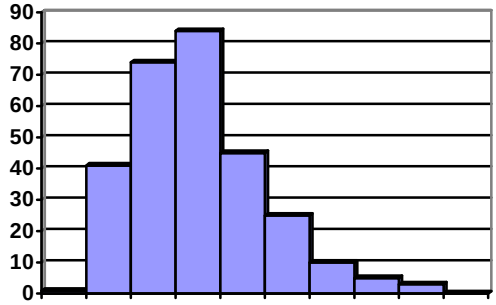

Figura 4.3: Exemplo de histograma assimétrico

- Despenhadeiro – o histograma possui um ou dois lados terminados de forma abrupta, aparentando faltar um pedaço do gráfico, como visto na Figura 4.4. Isto ocorre possivelmente devido a um ponto de corte nos dados, que espelha-se no formato do histograma.

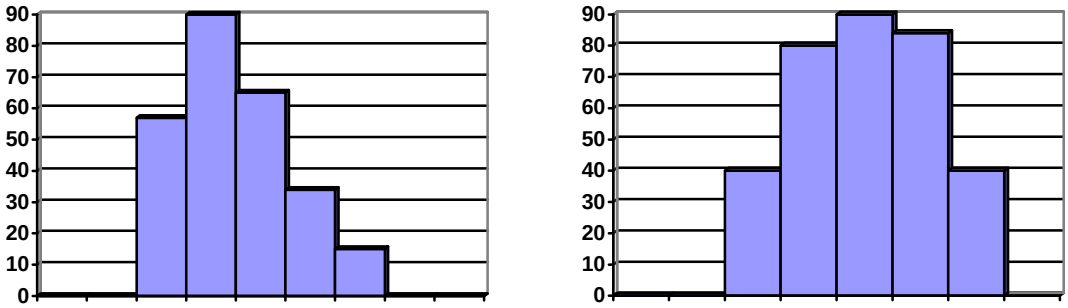

Figura 4.4: Exemplo de histograma despenhadeiro

- Dois picos – ocorre em situações em que são analisados dados diferentes, como por exemplo, a análise de dois tipos de matérias primas. A Figura 4.5 mostra um exemplo desse tipo.

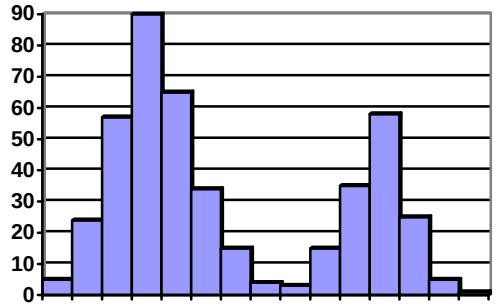

Figura 4.5: Exemplo de histograma de dois picos

- Platô – as classes centrais possuem frequências semelhantes, ocorre quando há mistura de distribuições com médias diferentes. A Figura 4.6 traz um exemplo desse tipo de histograma.

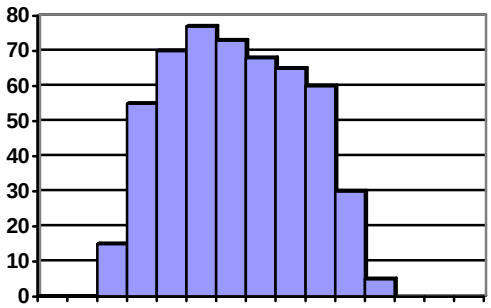

Figura 4.6: Exemplo de histograma platô

- Ilha isolada – algumas classes encontram-se separadas do restante das outras, formando um pequeno agrupamento, como ilustra a Figura 4.7. Isto ocorre devido a algum tipo de anomalia, tais como erros de medição, ou transcrição dos dados, falha no processo, podendo ser um indicativo de *outlier*.

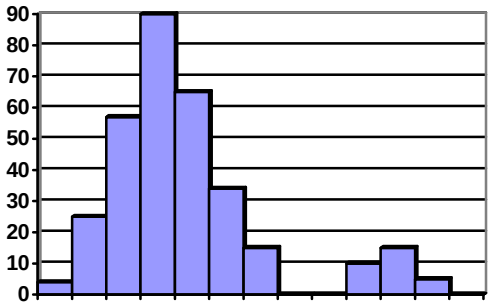

Figura 4.7: Exemplo de histograma de ilha isolada

a.III) Dimensionalidade

Possui dimensionalidade baixa, representando somente um atributo.

a.IV) Volume

De acordo com estudos de casos apresentados por (NIST / SEMATECH, 2006) e Myatt (2007; p.116), o histograma é aplicado para contagens amostrais de uma população. A princípio, o volume de registros manipulados pode ser grande, entretanto, a representação gráfica geralmente

é realizada sobre os valores numéricos obtidos de cálculos estatísticos, o que reduz o montante de registros a volume pequeno de dados.

a.V) Posição

Trata da distribuição de frequência para um único atributo, sendo este parâmetro, portanto, não relevante para esta técnica.

b) Diagrama de caixa (*Box plot*)

O diagrama de caixa apresenta uma visão geral da distribuição de frequência para um atributo, como ilustra a Figura 4.8, sendo possível detectar *outliers*, indicados por pontos localizados além dos valores mínimo ou máximo. Assim como no gráfico de dispersão, o diagrama de caixa possibilita a comparação entre os atributos, por meio do qual é possível visualizar:

- valor mínimo para o atributo;
- quartil inferior: representa 25% dos valores da amostra ordenada;
- mediana ou segundo quartil: é o valor até ao qual encontram-se 50% da amostra ordenada;
- quartil superior: encontram-se os valores mais elevados da amostra, abrangendo 75% da amostra ordenada;
- valor máximo para o atributo.

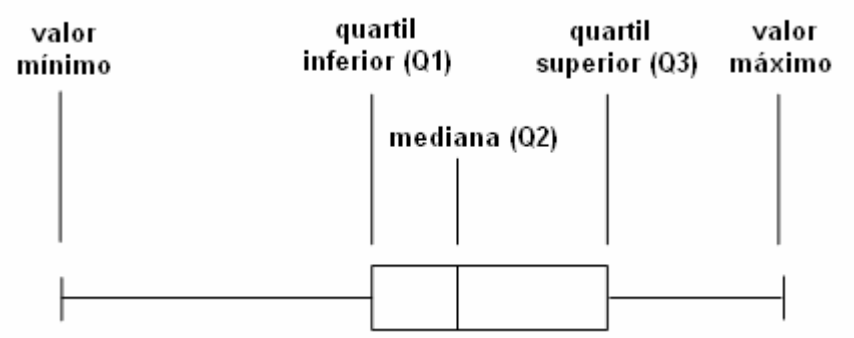

Figura 4.8: Formato do diagrama de caixa (adaptado de (Myatt, 2007))

b.I) Tipo de tarefa

Neste tipo de visualização, diferentes classes de dados podem ser comparadas quanto à localização (mediana, quantis) e variação da distribuição de frequência. *Outliers* podem ser detectados quando há dados representados para fora dos limites dos valores mínimo ou máximo do diagrama de caixa.

No exemplo disponível em (NIST / SEMATECH, 2006), ilustrado pela Figura 4.9, está a comparação de quatro máquinas (eixo *x*) quanto à sua produção de energia (eixo *y*). Pode-se observar que a máquina 3 apresenta o maior índice de energia, em torno de 72,5. Ainda, a máquina 4 apresenta menor variação na produção de energia, com cerca de 50% das leituras correspondendo a uma unidade de energia (NIST / SEMATECH, 2006).

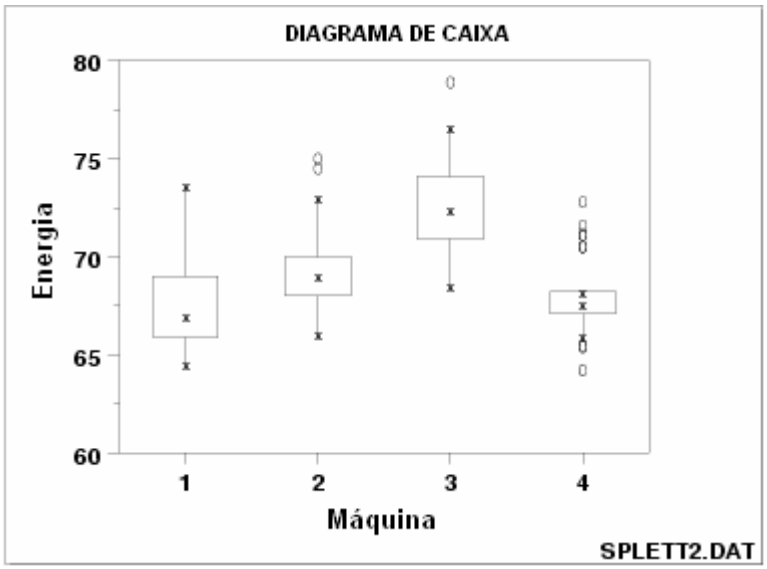

Figura 4.9: Diagrama de caixas para comparar quatro grupos de dados (*adaptado de (NIST / SEMATECH, 2006))*

b.II) Tipo de dado

O diagrama de caixa representa dados quantitativos contínuos (Myatt, 2007, p. 52).

b.III) Dimensionalidade

Assim como o histograma, o diagrama de caixa possui dimensionalidade baixa, representando somente um atributo. Porém, diferentes grupos para um mesmo atributo podem ser comparados entre si, como ilustrado na Figura 4.9.

b.IV) Volume

O diagrama de caixas representa sucintamente as distribuições de frequência de um atributo, obtidas por meio de cálculos que geralmente são feitos sobre um conjunto de dados pequeno, como mostra um estudo de caso apresentado em (NIST / SEMATECH, 2006), no qual é aplicado o diagrama de caixas para representar uma base de dados com 480 registros.

b.V) Posição

Como o diagrama de caixa é utilizado para a análise da distribuição de frequência de um atributo, este parâmetro não exerce influência alguma.

c) Gráfico de dispersão

O gráfico de dispersão é baseado no sistema de coordenadas cartesianas, no qual os valores dos atributos a serem observados são distribuídos para os eixos *x* e *y*. Este par de valores é observado quanto ao comportamento de um atributo em relação ao outro.

c.I) Tipo de tarefa

Por meio do gráfico de dispersão, é possível verificar se o par de atributos está correlacionado de modo positivo (*y* cresce quando *x* cresce), negativo (*y* decresce quando *x* cresce), como ilustra a Figura 4.10, ou correlação inexistente, como mostra a Figura 4.11.(a). Entretanto, a correlação detectada entre os atributos nem sempre significa relação de causa-efeito, em que aumentos sucessivos nos valores de um atributo acarretarão em aumentos sucessivos (ou diminuições sucessivas) no outro (NIST / SEMATECH, 2006). Também é possível a detecção de *outliers*, como mostra a Figura 4.11.(b), na qual pode-se observar um registro cujo valor é discrepante em relação aos demais representados na visualização.

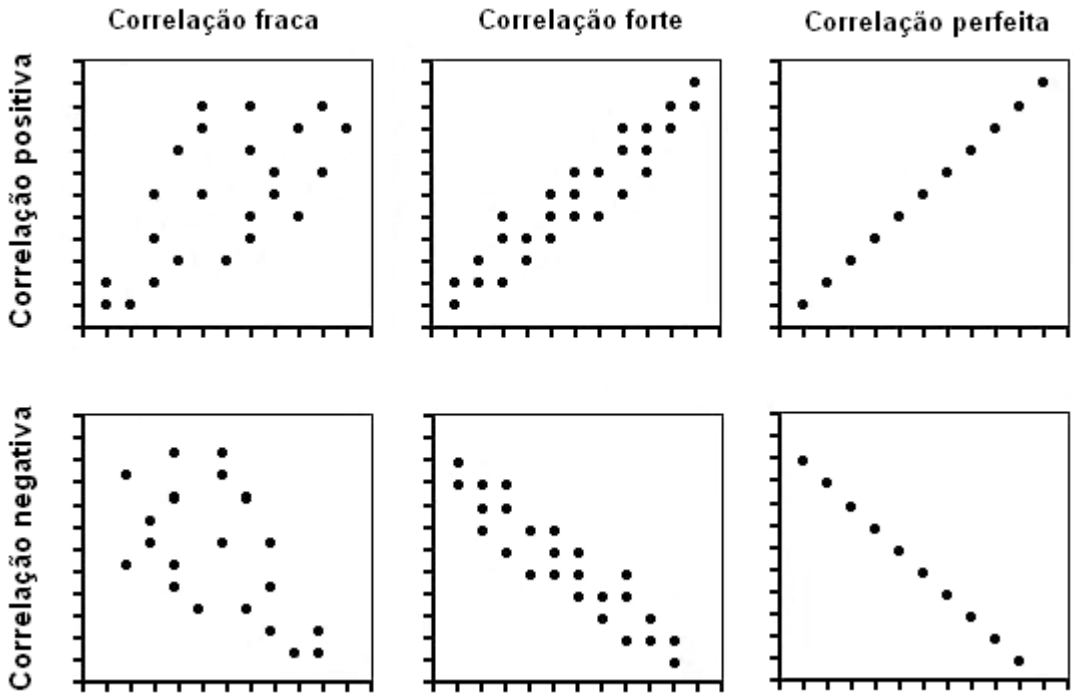

Figura 4.10: Correlação positiva e negativa entre dois atributos observada pelo gráfico de dispersão (Lugli, 2009)

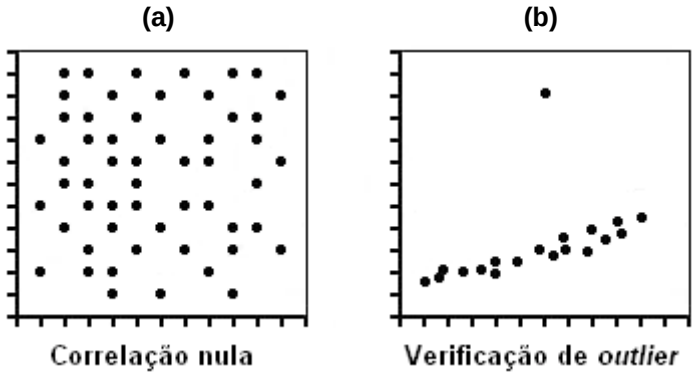

Figura 4.11: (a) Correlação nula entre dois atributos observada pelo gráfico de dispersão (Lugli, 2009); (b) Identificação de outlier (adaptado de (NIST / SEMATECH, 2006))

Vários gráficos de dispersão podem ser combinados para visualizar a correlação de mais de dois atributos, como é feita na técnica Matriz de Dispersão, em que são gerados todos os pares de combinações possíveis dos atributos.

c.II) Tipo de dado

Segundo Myatt (2007; p. 91) "geralmente é mais informativo explorar duas variáveis contínuas, correspondendo a uma razão ou intervalo de valores, utilizando a matriz de dispersão".

c.III) Dimensionalidade

Esta técnica é utilizada para a visualização da correlação entre dois atributos.

c.IV) Volume

Como cada ponto desenhado no gráfico representa um valor de dado, pelos exemplos citados em Myatt (2007) e (NIST / SEMATECH, 2006), pode-se observar que o volume de dados é pequeno. Myatt (2007, p. 43) afirma ainda que é possível representar centenas de observações utilizando o gráfico de dispersão.

c.V) Posição

Em (NIST / SEMATECH, 2006) é explicado que no eixo vertical (variável *y*), são geralmente atribuídos os valores resposta; e no eixo horizontal (variável *x*), são geralmente atribuídos os valores que supostamente possam estar relacionados à variável resposta. Dessa forma, pode-se

verificar pelo gráfico como os valores em *x* comportam-se em relação aos valores de *y*. Assim, a variação da posição não é determinante para a interpretação dos resultados representados, dado que a locação das variáveis é bem definida de acordo com a análise dos atributos envolvidos.

d) Gráfico de contorno

O gráfico de contorno representa o relacionamento entre três variáveis em uma visão bidimensional. Duas variáveis independentes são atribuídas aos eixos *x* e *y* do plano cartesiano e uma terceira variável *z* é utilizada para demarcar os níveis de contorno. O propósito do gráfico de contorno é de observar z em função das variáveis *x* e *y*. As regiões entre os contornos podem ser coloridas ou sombreadas para indicar seu grau de magnitude.

d.I) Tipo de tarefa

Uma das aplicações deste gráfico é a construção de mapas topográficos. A Figura 4.12 mostra um exemplo de utilização do gráfico de contorno para demonstrar o geóide da Terra, (forma que a Terra teria, se a superfície de nível médio das águas do mar se prolongasse através dos continentes), em que os dados estão em metros (fonte: http://www.mathworks.com/access/helpdesk/help/toolbox/map/contourfm.html, acesso em 01 de dezembro de 2009).

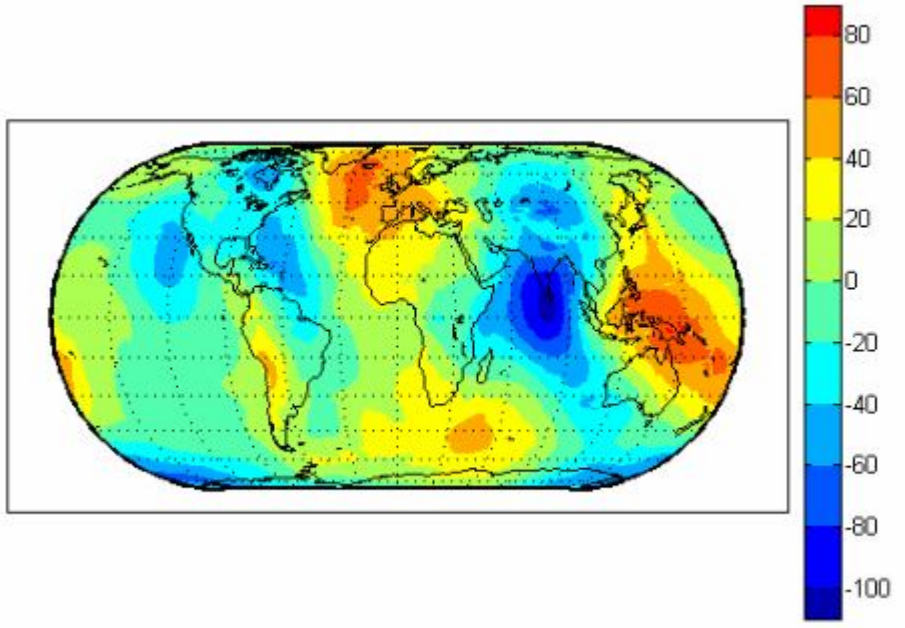

Figura 4.12: Exemplo de utilização do gráfico de contorno

d.II) Tipo de dado

O gráfico de contorno representa graficamente dados quantitativos discretos e contínuos, como mostram, por exemplo, trabalhos como os de Teles e Flôres (2007), Freitas e Jackix (2004) e Seidel e Rappaport (1992). As Figuras 4.13 e 4.14, a seguir, exibem a utilização do gráfico de contorno para explicar os dados utilizados ou obtidos nas pesquisas desses dois últimos trabalhos citados.

A Figura 4.13 refere-se à concentração dos elementos frutoligossacarídeo (FOS) e pectina (PEC) em um experimento envolvendo o estudo do suco misto de cenoura e laranja (Freitas e Jackix, 2004).

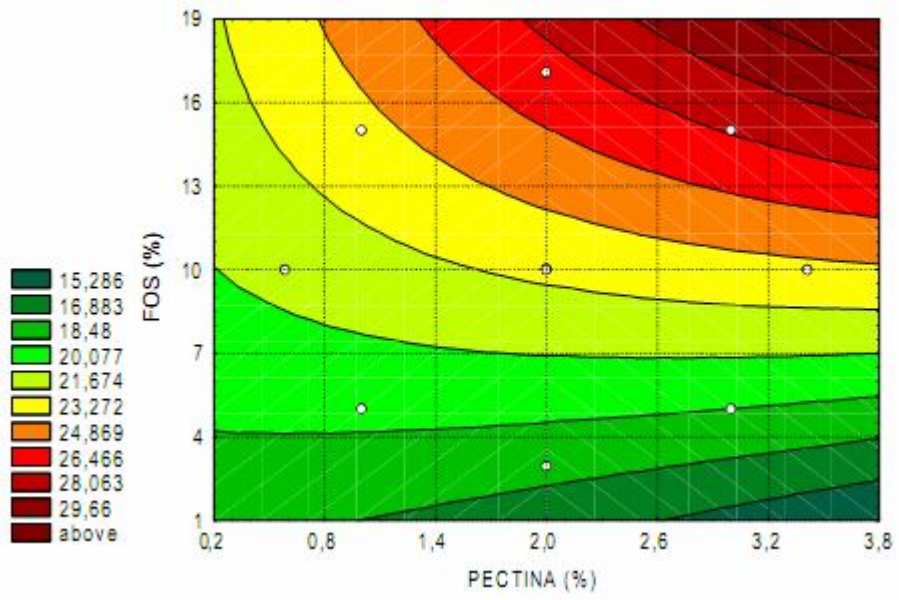

Figura 4.13: Exemplo de aplicação do gráfico de contorno (Freitas e Jackix, 2004)

Na Figura 4.14 tem-se a ilustração do resultado de um experimento sobre a transmissão e recepção de ondas de rádio dentro de um edifício, em que o gráfico de contorno mostra o alcance do transmissor utilizado (na figura, indicado por *Transmitter x*) em um dos andares do prédio (Seidel e Rappaport,1992).

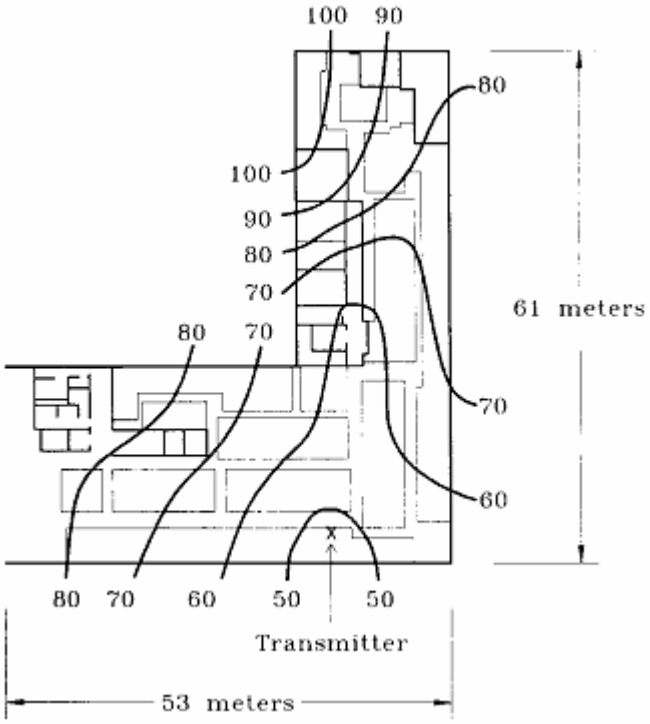

Figura 4.14: Exemplo de utilização do gráfico de contorno (Seidel e Rappaport, 1992)

d.III) Dimensionalidade

O gráfico de contorno ilustra o comportamento de um atributo em função de outros dois, representando, dessa forma, a correlação entre três atributos.

d.IV) Volume

O trabalho de Freitas e Jackix (2004) e Seidel e Rappaport (1992) reportam experimentos que, em geral, manipulam um volume pequeno de dados, o que permite o emprego da análise estatística e, consequentemente, o emprego de gráficos estatísticos.

d.V) Posição

Em definição dada em (NIST / SEMATECH, 2006), as variáveis independentes são atribuídas aos eixos *x* e *y*, e as linhas do contorno são desenhadas conforme uma terceira variável, *z*, que está em função de *x* e *y*. Da mesma forma que as outras técnicas da classe de gráficos 1D a 3D, o posicionamento dos atributos é pouco relevante devido à definição do objetivo de análise sobre as variáveis envolvidas.

e) Técnicas relacionadas

As demais técnicas, tais como gráfico de barras (Figura 4.15.a), gráfico de barras tridimensionais (Figura 4.15.b), gráfico de linhas (Figura 4.15.c), gráfico de pizza (Figura 4.15.d), gráfico de superfícies (Figura 4.15.e), gráfico de bolhas (Figura 4.15.f), entre outras, são amplamente discutidas em bibliografias da estatística, sendo usualmente utilizadas quando o objetivo é apresentar resultados, assim como as técnicas discutidas anteriormente.

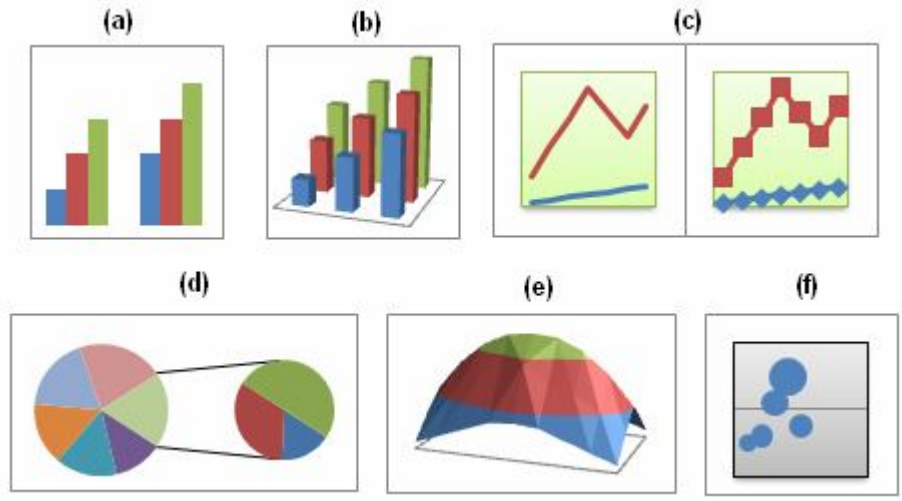

Figura 4.15: Técnicas da classe de gráficos 1D a 3D (fonte: <http://office.microsoft.com/pt-br/excel/HA010346071046.aspx?pid=CH062528081046> acessado em 14 de abril de 2010)

4.4.2. Técnicas iconográficas

As técnicas iconográficas utilizam uma entidade gráfica dotada de propriedades geométricas (forma, tamanho, posição e direção) e características de aparência (cor, textura, transparência) que são modificadas de acordo com o valor do atributo do dado que representam (Ward, 2002). Segundo Ward (2008), a vantagem das técnicas iconográficas sobre as técnicas geométricas (tais como coordenadas paralelas, matriz de dispersão) é a facilidade de percepção de padrões em um subconjunto de atributos. Entretanto, ainda segundo este autor, os ícones possuem suas limitações. Essas técnicas estão restritas quanto à precisão na representação dos valores dos dados e também quanto ao número de registros. Uma grande quantidade de dados tende a causar sobreposição de ícones ou a necessidade de diminuir o ícone para adequar-se ao tamanho da tela de exibição, dificultando a identificação de padrões. Assim, essas técnicas são recomendadas para a análise de um conjunto de dados não muito grande.

Walsum et al. (1996) identificam dois tipos de representação iconográfica:

- ícone de modelo-fixo – possui um número definido de parâmetros que modificam o

formato do ícone. Como exemplos, têm-se as figuras de aresta e faces de Chernoff;

- ícone de modelo-amorfo – não possui uma forma pré-definida, sendo totalmente determinada pelas propriedades do dado que representam, também não tem um número fixo de parâmetros. Como exemplo, a técnica *streamlines* para a visualização de fluxos de fluídos (Laramee et al., 2004).

Nesta seção, como exemplos de técnicas iconográficas, são discutidas as técnicas: (a) Faces de Chernoff; (b) *Star Glyphs* e; (c) Figura de Arestas. *Star Glyphs* pode representar um número maior de atributos quando comparada a Faces de Chernoff, mas ambas apresentam restrições ao representar grande volume de dados. A técnica Figura de Arestas, por outro lado, supre este aspecto. Estas e outras características dessas técnicas de visualização são apresentadas a seguir.

a) Faces de Chernoff

Observando que o ser humano é sensível ao reconhecimento de expressões faciais, Chernoff (1973) criou uma representação gráfica cujos atributos de dados são mapeados para as propriedades de uma face, tais como olhos, nariz e boca. A Figura 4.16 ilustra um exemplo de ícone cujas propriedades da face foram baseadas nos trabalhos de Chernoff (1973) e Flury e Riedwyl (1981). Na Figura 4.17 é ilustrado um exemplo de variações que as propriedades da face podem sofrer. A seguir são descritas as características principais dessa técnica, conforme os parâmetros.

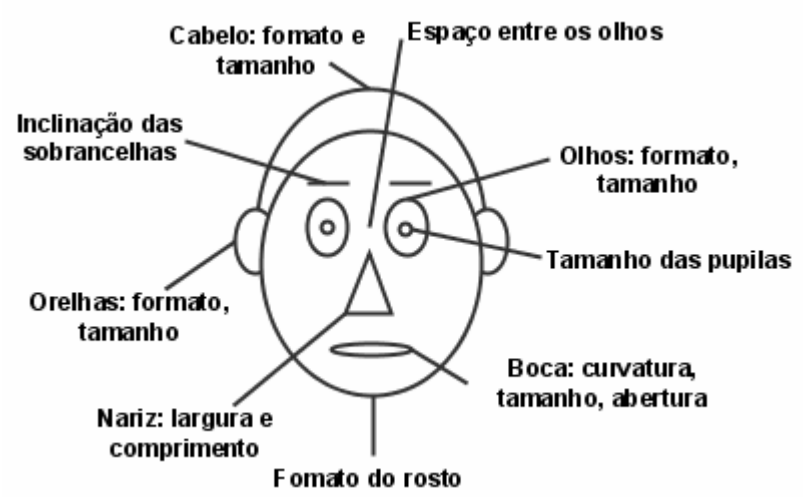

Figura 4.16: Propriedades da face de Chernoff

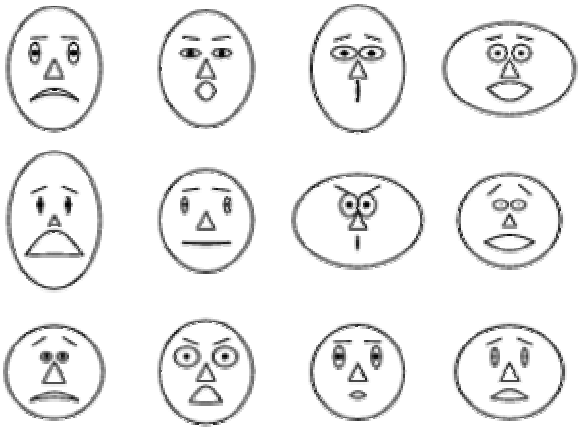

Figura 4.17: Exemplo de variações da face de Chenoff (Gonick e Smith, 1993)

a.I) Tipo de tarefa

As Faces de Chernoff podem ser utilizadas para identificar agrupamentos, por meio das semelhanças entre os ícones, detectar padrões de comportamentos nos dados e identificar *outliers*, pela verificação de um ícone com características discrepantes com o restante das faces (Bruckner, 1978; Saxena e Navaneetham, 1991). Entretanto, não é possível analisar o grau de relacionamento entre os atributos, pois os registros são apresentados separadamente por cada ícone (Rabelo, 2007).

No trabalho de Chernoff (1973), há dois exemplos de aplicação desta técnica. No primeiro, o especialista do domínio estava interessado em encontrar agrupamentos nos dados, e no segundo, outro analista desejava observar o comportamento dos dados em relação ao tempo por meio da sequência de faces apresentadas.

a.II) Tipo de dado

Os valores dos atributos são responsáveis pelas alterações das propriedades das faces, como posição dos olhos, tamanho do nariz, abertura da boca, dentre outras configurações permitidas. Desse modo, é conveniente que os valores dos atributos sejam numéricos, para que seja possível realizar o mapeamento para as propriedades da face.

Dentre trabalhos que apresentam a aplicação das faces de Chernoff estão: (Chernoff, 1973) no qual são representados dados de medições arqueológicas (dados quantitativos discretos); em (Bruckner, 1978) são utilizados dados financeiros (dados quantitativos contínuos); e (Flury e Riedwyl, 1981) e (Wyatt, 2008) apresentam mapeamento de dados populacionais (dados quantitativos contínuos) para as faces de Chernoff.

a.III) Dimensionalidade

Chernoff (1973) afirma que com esta técnica é possível mapear até 18 dimensões. Bruckner (1978) afirma que as faces de Chernoff são viáveis para representar até 15 atributos, pois acima deste valor a visualização é dificultada.

a.IV) Volume

Segundo Bruckner (1978), um grande volume de dado pode interferir na qualidade da visualização, principalmente quando vários ícones precisam ser exibidos simultaneamente, sendo necessária a redução do tamanho das faces para adequar-se ao espaço na tela. Isto compromete a visualização de detalhes mais sutis do ícone, como a inclinação da sobrancelha de uma face (Nascimento e Ferreira, 2005).

a.V) Posição

Devido ao fato do ícone seguir uma estrutura fixa, pois corresponde a uma face humana, a posição dos atributos não é uma questão relevante para essa técnica de visualização. Tanto que a ordem de distribuição dos atributos para as propriedades da face não é um alvo abordado nos trabalhos pesquisados. Entretanto, Chernoff (1973, p. 366) já questionava se algumas características da face sobressaíam-se sobre as demais. Segundo relatos sobre o experimento desse autor, o especialista que analisara dados de um dos exemplos declarou que os olhos possuem maior representatividade; o segundo especialista entrevistado, opinou que o formato da cabeça era a principal característica para observar os dados.

Em uma análise apresentada por Morris et al. (2000), é afirmado que o tamanho dos olhos e inclinação da sobrancelha são as propriedades mais perceptíveis à primeira vista (no experimento demonstrado por esses autores, a forma da cabeça manteve-se constante). Rabelo (2007) e Lee et al. (2003) afirmam que a facilidade de interpretação das faces de Chernoff depende de quão habituado o usuário está com o mapeamento dos atributos dos dados para as características do ícone.

b) *Star Glyphs*

Nesta técnica, *n* atributos do dado são mapeados para *n* linhas que partem de um ponto em comum, cujos tamanhos variam conforme o valor do atributo. As linhas são uniformemente

separadas por ângulos e as extremidades são unidas por um segmento de reta, formando um polígono.

b.I) Tipo de tarefa

Considerando o formato de cada estrela é possível detectar agrupamentos pela identificação de elementos com formatos semelhantes, detectar regras e padrões. *Outliers* podem ser identificados por ícones que apresentam formatos fora do padrão comum exibido (NIST/SEMATECH, 2006).

b.II) Tipo de dado

Geralmente a técnica S*tar Glyphs* é utilizada para visualizar dados quantitativos, embora outros tipos de dados também possam ser representados, porém, precisam ser transformados antes de serem mapeados, como é o caso de valores negativos (Klippel et al., 2009).

b.III) Dimensionalidade

A vantagem desta técnica em relação à face de Chernoff é a capacidade de representar um número maior de dimensões, entretanto, uma grande quantidade de atributos tende a formar borrões (Rabelo, 2007), como pode ser observado na Figura 4.18.

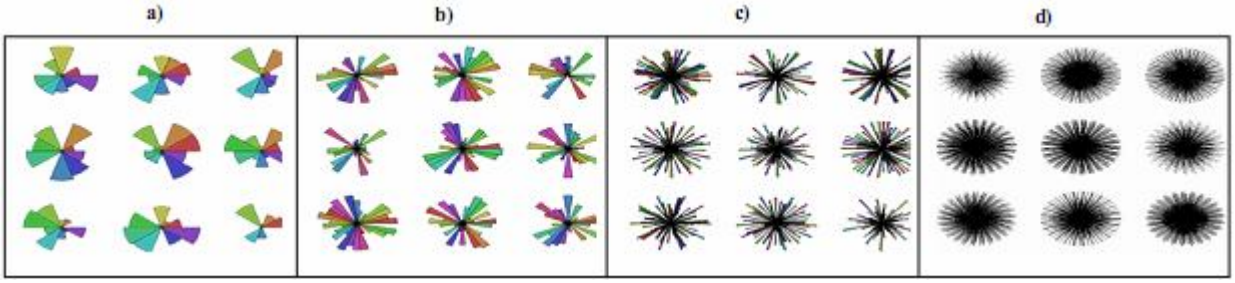

Figura 4.18: a) 10 atributos, b) 30 atributos, c) 80 atributos, d) 500 atributos (Rabelo, 2007)

b.IV) Volume

Assim como na face de Chernoff, cada ícone representa um registro de dado posicionado num espaço bidimensional. Desse modo, para um volume de dados muito grande, a visualização de forma geral pode ficar comprometida devido à diminuição dos ícones para ajustarem-se na tela (Ward, 2002). Em avaliação das técnicas de visualização feita por Rabelo (2007), o autor afirma que essa técnica possui baixo suporte a grandes quantidades de dados.

b.V) Posição

Peng et al. (2004), e Klippel et al. (2009) discutem a ordem de atribuição dos atributos para os raios, que alteram o formato dos ícones. Segundo Peng et al. (2004), o formato mais convexo, denominado *teardrop*, é o mais fácil para distinguir agrupamentos entre os dados. Entretanto, Klippel et al. (2009) afirmam que o formato que apresenta mais concavidades também é significativo na detecção de agrupamentos.

Contudo, uma vez definida a ordem de mapeamento dos atributos para as linhas da *Star Glyph*, significa que o posicionamento dos atributos para uma estrela é o mesmo em relação a qualquer registro representado no gráfico, anulando desse modo a influência do parâmetro posição dos atributos na análise dos dados por meio dessa técnica.

c) Figura de Aresta (S*tick Figure*)

O ícone é formado por segmentos de linha conectados, sendo um deles denominado corpo do ícone e o restante dos segmentos denominados limbos. Dois atributos são mapeados para as dimensões da tela e os demais são mapeados para ângulos, comprimentos e cores dos segmentos das figuras de arestas. A Figura 4.19 mostra uma configuração com 5 segmentos e suas possíveis variações.

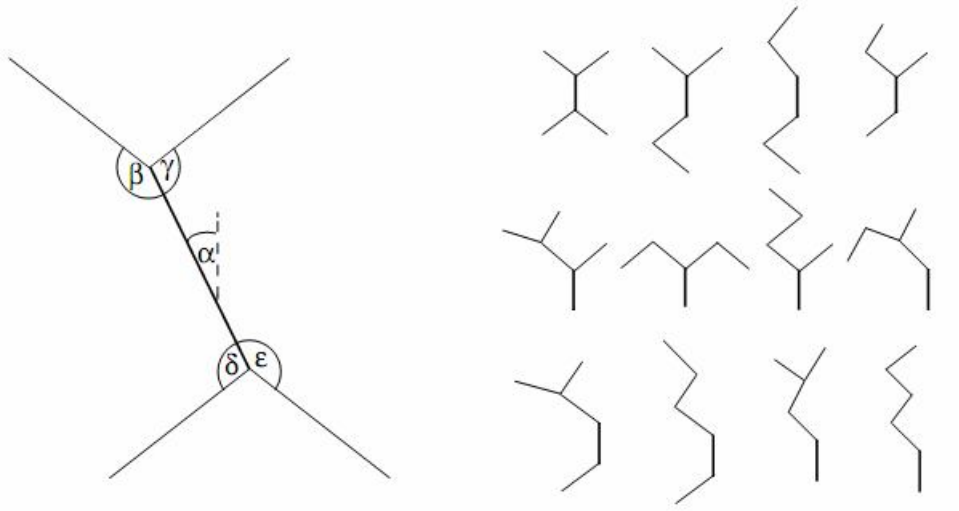

Figura 4.19: Figura de aresta e os membros que podem ser derivados (Keim, 1997)

c.I) Tipo de tarefa

Keim e Kriegel (1996) confirmam que esta técnica é útil na percepção de agrupamentos, sendo mais apropriada para dados que possuem semântica bi ou tridimensional. A percepção de agrupamentos se dá pela posição e pela estrutura formada de cada ícone, que resultam em

texturas facilmente identificáveis na imagem. A Figura 4.20.(a) ilustra um exemplo de textura formada pelo arranjo dos ícones gerados por cinco imagens de satélite da região dos Grandes Lagos (Figura 4.20.(b)), no qual é visto, especificamente, uma parte do oeste do Lago Ontário, a extrema ponta do Lago Erie e uma parte do leste do Lago Huron (Pickett e Grinstein, 1988; Grinstein et al., 2001).

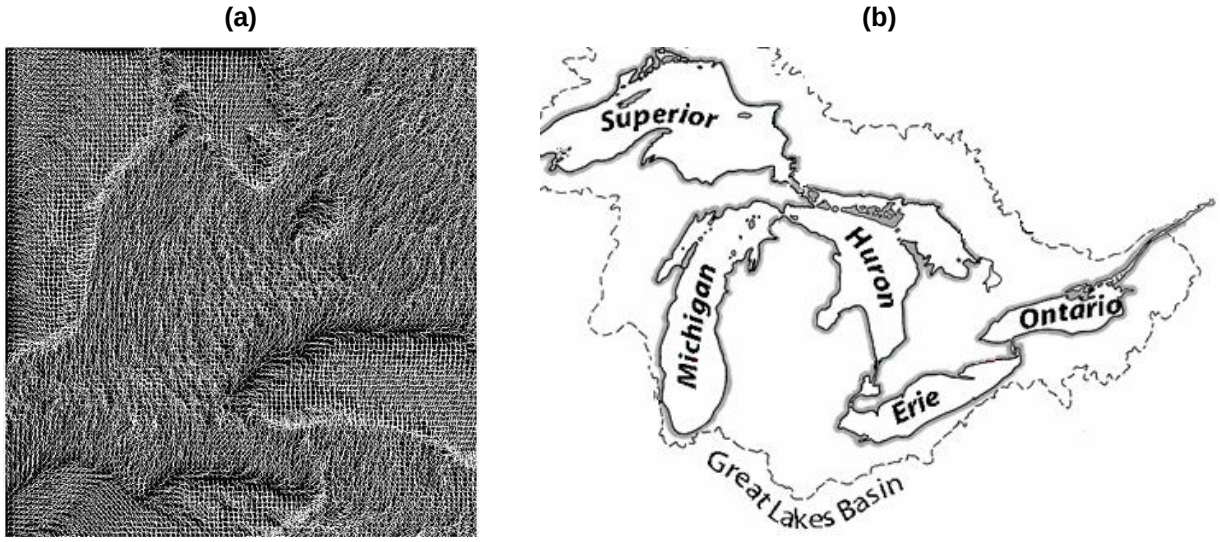

Figura 4.20: (a) Exemplo de textura formada pelas figuras de arestas; (b) Mapa dos Grandes Lagos da qual uma parte desta está mapeada em (a) (Pickett e Grinstein, 1988; Grinstein et al., 2001)

c.II) Tipo de dado

Pickett et al. (1988) afirmam que essa técnica é apropriada para analisar, principalmente, duas classes de dados. A primeira engloba dados físicos[4], que descrevem objetos do mundo real como, por exemplo, dados provenientes de satélites ou sensores sísmicos, sobre os quais a técnica pode ser aplicada para o estudo do comportamento de fenômenos naturais; dados médicos sobre imagens de ressonância magnética; dados sobre o fluxo dos fluídos, estudado na Física. A segunda classe refere-se a dados multidimensionais com propriedades temporais e/ou espaciais, características típicas de dados estatísticos, como dados de censo populacional ou de estudos epidemiológicos.

[4] Alguns autores (Shneiderman, 1996; Oliveira e Levkowitz, 2003; Keim 2008) atribuem o estudo desse tipo de dado à Visualização Científica, uma área à parte da Visualização da Informação. Entretanto, Rhyne (2003) argumenta que não há fronteiras nítidas entre essas áreas, sendo desnecessária essa distinção.

c.III) Dimensionalidade

Segundo Rabelo (2007), esta técnica pode representar dados com alta dimensionalidade, suportando um número de 80 atributos, que ainda permite uma distinção entre os elementos representados.

c.IV) Volume

De acordo Keim e Kriegel (1996) e Pickett e Grinstein (1988), a Figura de Aresta permite a visualização de grande quantidade de dados, quando comparada às demais técnicas iconográficas.

c.V) Posição

Segundo Pickett e Grinstein (1988, p. 516), não há uma regra para mapear as dimensões do dado para as propriedades do ícone, pois o mapeamento para uma família de arestas pode revelar-se mais informativa que outra. Segundo esses autores, este é o espírito da exploração visual: visualizar os dados de diferentes modos, até encontrar um padrão novo e interessante.

4.4.3. Técnicas geométricas

Dados multidimensionais são mapeados para um plano bidimensional, em que todos os atributos podem ser vistos simultaneamente, diferentemente das técnicas iconográficas, em que determinadas propriedades dos ícones podem dar mais destaques a determinados atributos. Segundo Oliveira e Levkowitz (2003), as técnicas geométricas podem manipular grande quantidade de dados, desde que sejam associadas outras técnicas de interação, para obter maior nitidez sobre a visualização. Para exemplificar essa classe, são descritas a seguir as técnicas: (a) Matriz de Dispersão e (b) Coordenadas Paralelas.

a) Matriz de Dispersão

Como o próprio nome sugere, constitui-se de uma matriz de gráficos de dispersão (visto na Seção 4.4.1.c) que permite analisar dados multidimensionais.

a.I) Tipo de tarefa

A matriz de dispersão exibe as possíveis combinações dos pares de atributos dos dados, o que permite observar as correlações de cada atributo com os demais (Grinstein et al., 2001). Cores e símbolos podem ser utilizados para distinguir agrupamentos (Carr et al., 1987).

a.II) Tipo de dado

Por ser composta de um conjunto de gráficos de dispersão organizados em uma matriz, esta técnica também contempla os mesmos tipos de dados: quantitativos contínuos.

a.III) Dimensionalidade

A alta dimensionalidade também prejudica a visualização geral dos dados, pois diminui a área de cada gráfico de dispersão. Segundo Rabelo (2007), a matriz de dispersão pode exibir cerca de 15 dimensões, considerando símbolos e cores, sem gerar borrões na visualização.

a.IV) Volume

A representação de grande volume de dados pode comprometer a eficiência da visualização, devido à sobreposição dos pontos (Carr et al., 1987).

a.V) Posição

Segue a mesma ideia dos gráficos de dispersão, sendo constituída por um conjunto destes, no qual a disposição dos atributos é predeterminada pelo objetivo do analista em verificar a correlação entre determinados atributos.

b) Coordenadas paralelas

Proposta inicialmente por Inselberg (1985; 2008), as coordenadas paralelas (Figura 4.21) mapeiam dados multidimensionais para um plano bi-dimensional, cuja representação se dá por meio de eixos verticais paralelos sobre os quais são atribuídas as dimensões dos dados, com seus valores mínimos e máximos distribuídos, respectivamente, para os limites inferiores e superiores do eixo, que podem ser normalizados, se necessário, seguindo a média, mediana, um caso específico ou um valor específico (Theus, 2008). Os eixos são interceptados por linhas poligonais que tocam nos pontos correspondentes ao valor do atributo, sendo cada linha correspondente a um registro.

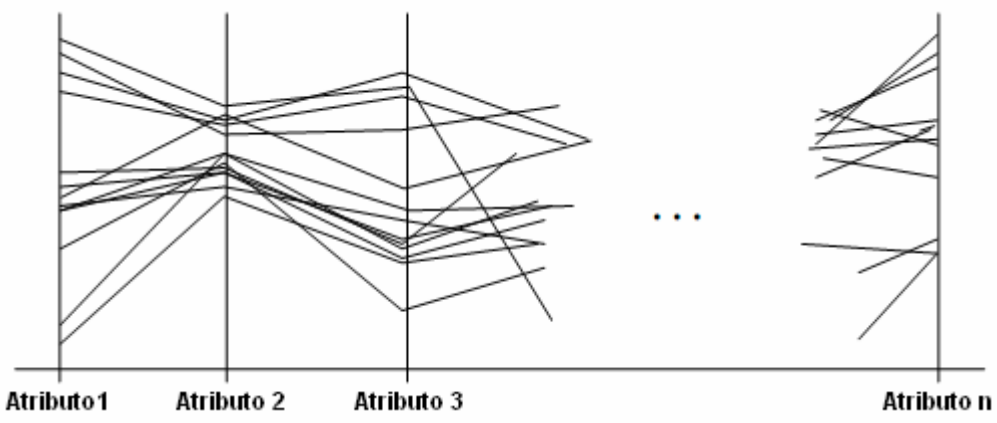

Figura 4.21: Coordenadas paralelas (adaptado de Keim, 1997)

b.I) Tipo de tarefa

De acordo com Theus (2008), além da visão geral, as coordenadas paralelas servem para observar o comportamento de um caso específico (que pode ser um único dado ou um subconjunto) por meio de seu destaque com o uso de cores ou saturação (grau de transparência com que é desenhada a linha) e fazer comparações com o restante dos dados. A Figura 4.22 é um exemplo que ilustra o destaque para um subconjunto de carros que estão entre os mais potentes (visto pelo atributo *Horsepower*) e mais caros (observado pelo atributo *Dealer Cost* – preço na revendedora).

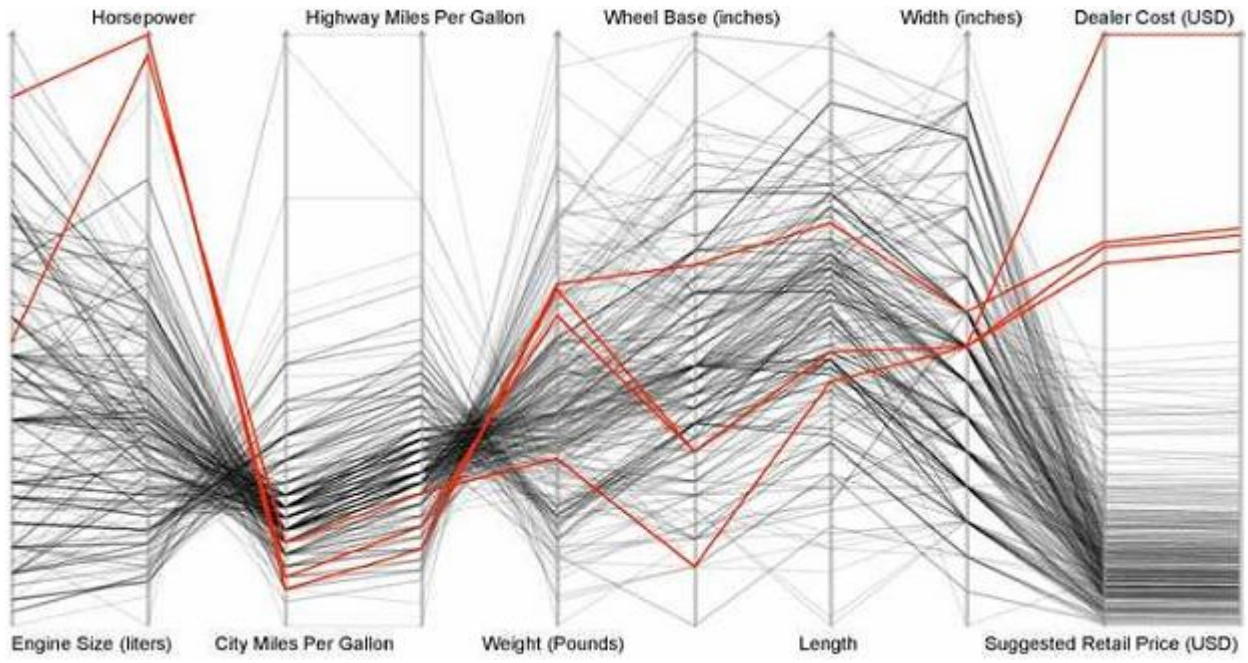

Figura 4.22: Exemplo de utilização de coordenadas paralelas para destacar um subconjunto de dados (Theus, 2008)

Por meio desta técnica pode-se observar a existência de correlações positivas (*y* cresce quando *x* cresce) marcadas por linhas com inclinação semelhante, e correlações negativas (*y* decresce quando *x* cresce), detectadas pela formação de cruzamento entre os eixos. As Figuras 4.23 e 4.24 mostram um exemplo de comportamento das variáveis Var1, Var2, Var3, Var4, na qual primeiramente são visualizadas pelas coordenadas paralelas (Figura 4.23) e depois pela matriz de dispersão (Figura 4.24). Pode-se observar que Var1-Var2 não apresentam correlação; Var2-Var3 apresentam forte correlação positiva; Var3-Var4 apresentam forte correlação negativa.

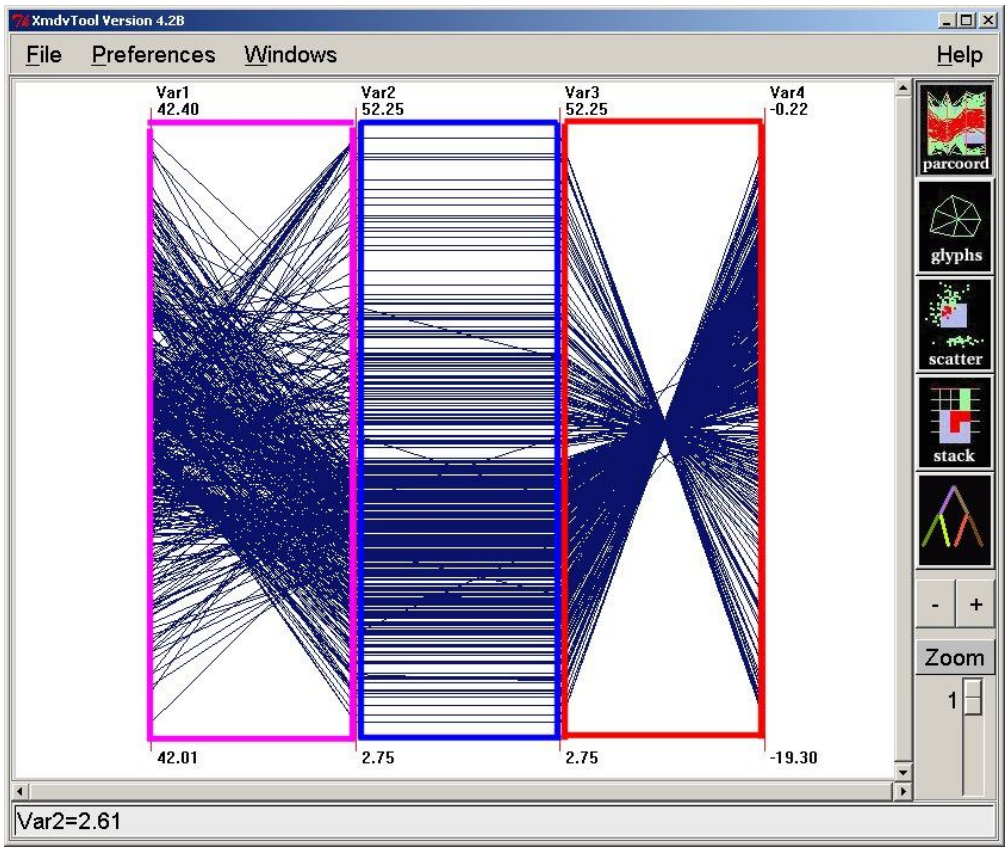

Figura 4.23: Correlação entre variáveis Var1, Var2, Var3, Var4, *vistas por meio da ferramenta Xmdv utilizando coordenadas paralelas (EVL, 2009).*

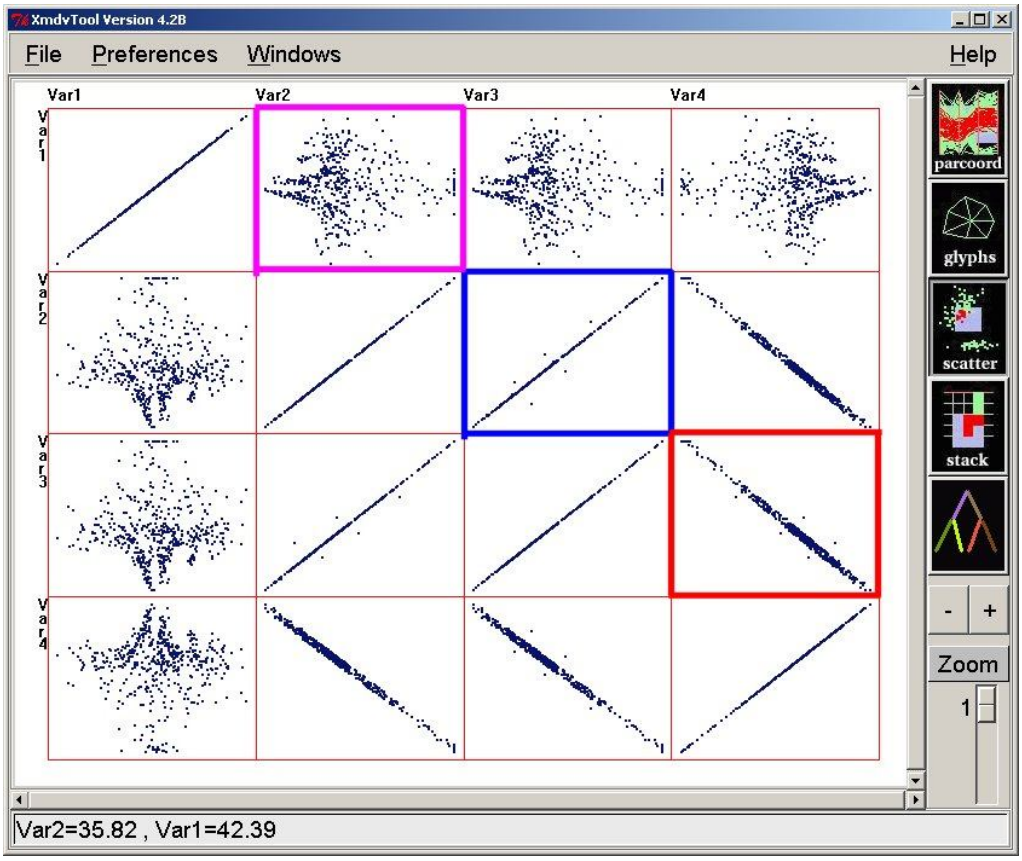

Figura 4.24: Correlação entre variáveis Var1, Var2, Var3, Var4, *vistas por meio da ferramenta Xmdv utilizando matriz de dispersão. (EVL, 2009).*

b.II) Tipo de dado

Esta técnica permite a visualização de vários atributos ao mesmo tempo, que podem ser tanto qualitativos quanto quantitativos (Rabelo, 2007), sendo especialmente útil para representar dados que possuem um tipo de ordenação, como ordem cronológica, ou cujos atributos compartilham uma escala em comum (Theus, 2008).

b.III) Dimensionalidade

As coordenadas paralelas comportam alta dimensionalidade dos dados. Inselberg (2008, p.663) afirma já ter trabalhado com uma base de dados com 800 atributos e 10.000 registros. Mas a visualização pode ficar comprometida, pois os eixos verticais tendem a ficar muito próximo por causa do espaço de exibição do gráfico.

b.IV) Volume

Como cada linha poligonal representa um item entre os dados, a representação de um grande volume pode gerar "borrões" devido à sobreposição dos itens, prejudicando a detecção de agrupamentos ou padrões. Segundo Keim e Kriegel (1996), devido a esta limitação, esta técnica é adequada para analisar um conjunto de dados de volume médio, com grande dimensionalidade, sendo especialmente útil para obter uma visão geral dos dados e descobrir correlações entre os atributos.

b.V) Posição

A ordem de disposição dos eixos é crucial na interpretação de padrões contidos nos relacionamentos entre atributos (Theus, 2008). Portanto, é importante a permutação das posições dos eixos durante a exploração dos dados (Inselberg, 2008). Segundo Wegman (1990), o número mínimo de diferentes ordenações para visualizar todas as adjacências entre os atributos é da ordem de $[(n+1)/2]$ para dados com n atributos.

4.4.4. Técnicas orientadas a pixel

Nesta seção são apresentadas as técnicas: (a) *query-independent* e (b) *query-dependent*, descritas em maiores detalhes nos trabalhos de Keim e Kriegel (1996) e Keim (2000). O que basicamente diferencia as duas técnicas discutidas é a forma de organização dos pixels no gráfico. Os demais parâmetros possuem basicamente os mesmos conceitos preservados para essas duas técnicas. Assim, os parâmetros: (I) tipo de tarefa, (II) tipo de dado, (III) dimensionalidade e (IV) volume são apresentados de forma geral para essa classe, sendo somente o parâmetro posição descrito em particular para cada técnica.

Nas técnicas orientadas a pixel, o posicionamento dos atributos interfere na análise da correlação entre as dimensões, por isso, para verificar o relacionamento entre os atributos, convém dispô-los próximos na visualização. Uma particularidade tratada nessa classe de técnicas é o algoritmo de posicionamento dos pixels na geração da visualização. Por esta razão, o enfoque dado para o parâmetro posição, neste caso, é dado a esses algoritmos, como resume a Tabela 4.11. Dessa forma, segundo Keim (2000; p.03), a disposição dos elementos pode obedecer à técnica dirigida a dados (*query-independent technique*) ou dirigida a resultados (*query-dependent technique*).

A primeira técnica é apropriada para detectar padrões em dados temporais, que possuem uma ordem natural de acordo com um atributo (dias, semanas, meses, anos) ao qual pode-se aplicar algum tipo de agrupamento. A segunda pode ser utilizada para verificar o resultado de uma consulta sobre os dados. Neste caso, em vez do valor do atributo, os pixels representam a distância dos elementos em relação ao resultado da consulta. Assim, os pixels correspondentes aos elementos que satisfazem a consulta são centralizados no meio da janela e, conforme o valor dos elementos distancia-se do resultado, os respectivos pixels vão sendo dispostos em direção à extremidade da janela, ordenados segundo o cálculo de uma distância global.

Tabela 4.11: Tipos de ordenação dos pixels (Keim, 1997)

Técnica	**Formato de organização dos pixels**
Dirigido a dados *Query-independent technique*	Arranjos simples
	Curvas de preenchimento (S*pace-filling curves*)
	Técnica do padrão recursivo (*Recursive Pattern*)
Dirigido a resultados *Query-dependent technique*	Técnica Espiral (*Spiral technique*)
	Técnica dos Eixos (*Axes technique*)
	Técnica Segmentos circulares (*Circle segment*)

I) Tipo de tarefa

As técnicas orientadas a pixel servem para visualizar relações entre os atributos. Em estudo realizado por Keim e Kriegel (1996), em que é feita uma comparação entre esta técnica com figuras de arestas e coordenadas paralelas, os autores afirmam que as técnicas orientadas a pixel possuem a vantagem de representar a sobreposição de dados, fato que ocorre quando o volume de dados é muito grande.

O formato retangular, como apresentado na Figura 2.12, pode dificultar a detecção de relacionamentos quando os dados possuem grande quantidade de atributos (dimensões), devido à distância relativa entre as janelas correspondentes a cada dimensão. Outra alternativa é o formato circular, adotado na técnica *Circle Segments*, ilustrada na Figura 4.25.

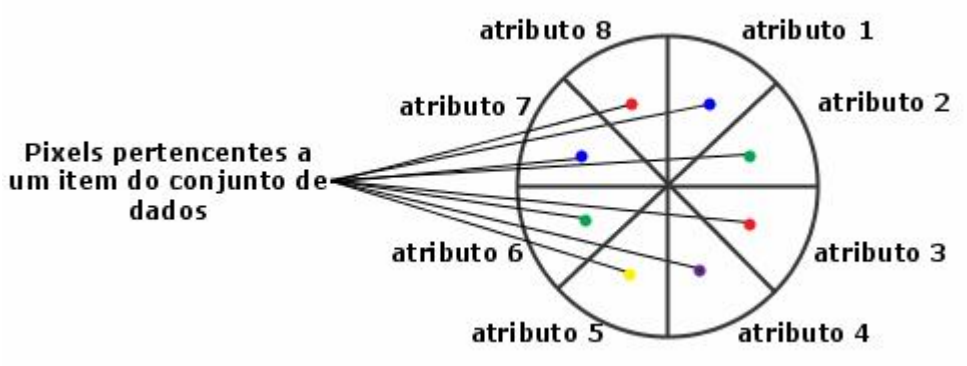

Figura 4.25: Aplicação alternativa da técnica orientada a pixel (adaptado de (Keim, 2000))

II) Tipo de dado

Segundo Keim e Kriegel (1996) e Keim (2000), as técnicas orientadas a pixel são adequadas para dados multidimensionais, em que a tela de exibição é dividida em *m* janelas, uma para cada atributo. Observando os exemplos apresentados por Keim (1996), essa técnica é destinada para visualizar dados quantitativos contínuos.

A cada pixel é atribuída uma cor para designar o elemento do dado. As cores possuem a vantagem de identificar maiores pontos de percepção em uma imagem (denominado por Keim como *just noticeable difference*). Keim (2000) adverte que o mapeamento deve ser feito cuidadosamente de modo que a visualização das cores fique evidente ao leitor. Unwin (2008) apresenta considerações a serem tomadas sobre a construção de gráficos, dentre as quais a questão da cor também é abordada, conforme discutido na Seção 2.3.2.

De acordo com Keim (2001), esta técnica não é apropriada para dados categóricos, sendo empregada experimentalmente sobre dados quantitativos temporais e dados quantitativos contínuos (Keim, 2000; 1996). Em (Keim, 2000), há uma demonstração do uso das técnicas de visualização para representar uma base de dados de valores financeiros diários correspondentes ao período de setembro de 1987 a fevereiro de 1995. Em outro trabalho apresentado (Keim, 1996), as técnicas orientadas a pixel foram utilizadas para representar dados de moléculas biológicas mapeadas em relação aos ângulos mínimos e máximos que constituiam seu formato.

Como mais um exemplo de aplicação das técnicas orientadas a pixel, pode-se citar o trabalho de Shimabukuro (2004), que utiliza as técnicas orientadas a pixel na implementação de uma ferramenta para visualizar dados de precipitação pluviométrica, medidas diariamente ao longo dos anos, em estações de coleta. Isto significa a existência de dados com características temporais, devido à periodicidade de tempo, e dados espaciais, representando a localização geográfica de cada estação (Shimabukuro, 2004; p.48). Assim, o mapeamento consiste da atribuição de uma cor de pixel para representar os dados de precipitação de chuva, sendo a

organização dos pixels feita de acordo com uma escala temporal (dia, ano, hora, semestre), seguindo um arranjo simples, que pode ser associado a uma escala espacial permitindo, por exemplo, observar os dados de uma estação de coleta.

III) Dimensionalidade

As técnicas orientadas a pixel podem exibir dados com alta dimensionalidade (Keim, 2000). Dependendo da quantidade de atributos, o analista pode optar por diferentes formatos de exibição, tais como janelas retangulares ou círculo.

IV) Volume

Essa classe de técnica permite a visualização de grande quantidade de dados. Segundo Keim (2000), aproximadamente cerca de 1.000.000 de valores, para uma tela de computador com resolução 1280x1024.

V.a) Posição dos pixels na técnica *query-independent*

Quando trata-se da técnica dirigida a dados (*query-independent*), a organização dos pixels pode ser feita por direções simples da esquerda para direita, seguindo linhas horizontais ou verticais, denominadas por Keim (1997) como *line-by-line* e *column-by-column*, respectivamente. Entretanto, segundo Keim (1996), este tipo de arranjo não oferece bons resultados, tendo como outras alternativas as curvas de Peano-Hilbert ou de Morton (*space-filling curves*), que possibilitam um posicionamento mais próximo entre as variáveis observadas. Porém, o traçado dessas curvas pode oferecer dificuldades de interpretação por parte do usuário, devido à complexidade do desenho que formam. Assim, foi proposta uma terceira técnica para arranjar os pixels, denominada *Recursive Pattern* (Keim, 2000). A Figura 4.26 ilustra os diferentes arranjos mencionados.

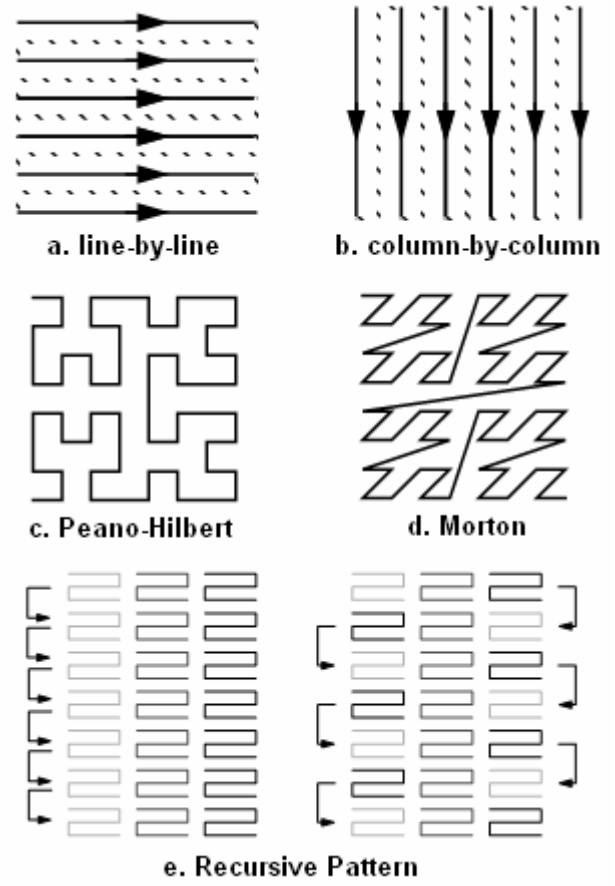

Figura 4.26: Formatos de arranjos dos pixels (adaptado de (Keim, 1996))

V.b) Posição dos pixels na técnica *query-dependent*

Quando é aplicada a técnica dirigida a resultados (*query-dependent*), a organização dos pixels pode seguir um formato espiral ou os pixels podem ser distribuídos entre eixos divisórios, que dividem a janela em quatro partes. Neste caso, dois atributos são alocados nos eixos e os valores das distâncias de cada elemento em relação ao resultado são posicionados conforme seus valores sejam positivos ou negativos. Ainda, os pixels podem ser ordenados em um formato circular.

A Figura 4.27 ilustra a ideia da organização em espiral. Na Figura 4.28 pode ser visto o resultado da visualização de um dado com cinco atributos, sendo o cálculo da distância global entre estes atributos apresentado na primeira janela à esquerda. Na Figura 4.29 é ilustrada a ideia da organização em eixos cujo exemplo é mostrado na Figura 4.30. Os pixels também podem ser organizados sobre um círculo, posicionados a partir do centro, conforme ilustra a Figura 4.31.

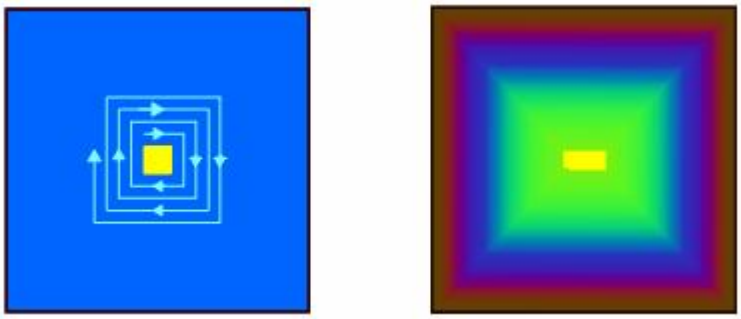

Figura 4.27: Técnica Espiral e exemplo do resultado da disposição global da distância dos atributos representados (Keim, 1997)

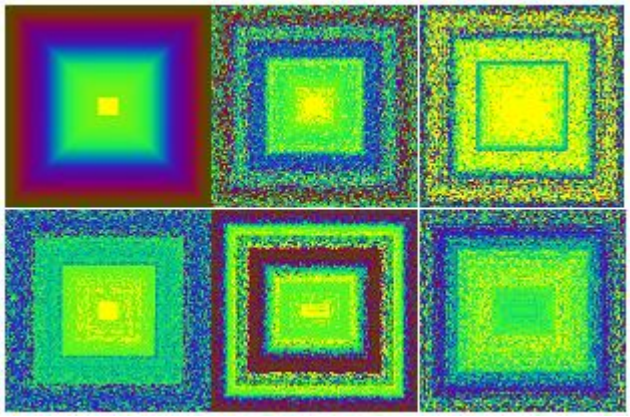

Figura 4.28: Exemplo de aplicação da técnica Espiral sobre um dado com cinco atributos (Keim, 1997)

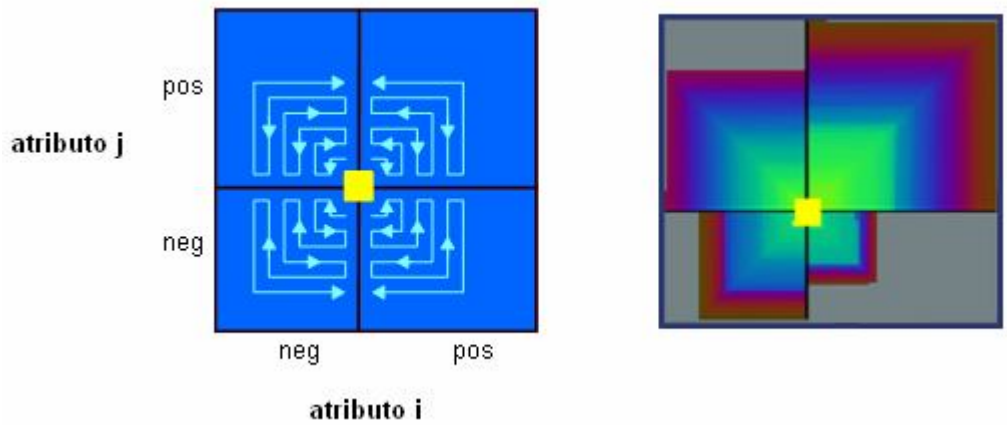

Figura 4.29: Técnica de Eixos e exemplo do resultado da disposição global da distância dos atributos representados (Keim, 1997)

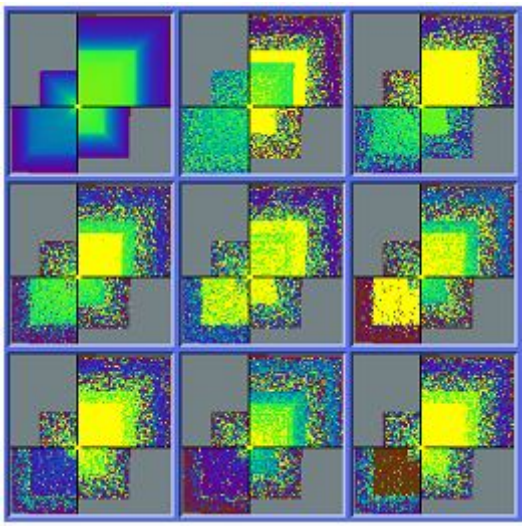

Figura 4.30: Exemplo de aplicação da técnica de Eixos sobre um dado com oito atributos (Keim, 1997)

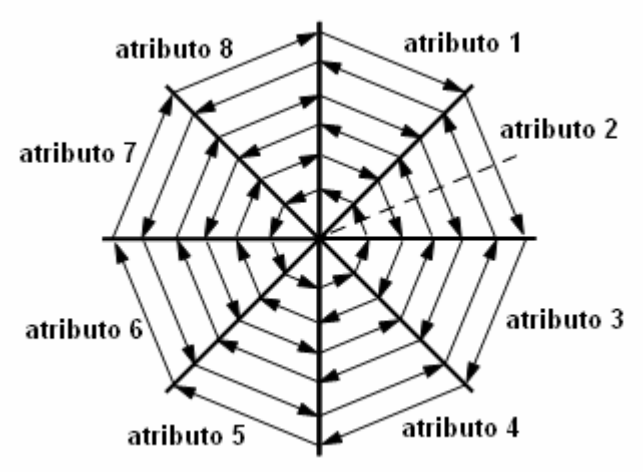

Figura 4.31: Técnica de Segmentos Circulares para dado com 8 atributos (adaptado de (Keim, 2000))

4.4.5. Grafos e técnicas hierárquicas

São próprias para dados que possuem naturalmente algum tipo de ligação entre si. Dessa forma, os dados são representados por uma estrutura composta por níveis e subníveis. Um exemplo típico é a estrutura de diretórios e subdiretórios de um sistema de arquivos, visualizada por meio de uma estrutura de árvore. Árvore é um tipo especial de grafo, composto por vértices, que representam os dados, ligados por arestas, que representam o relacionamento. Segundo Urbanek (2008), as técnicas hierárquicas que utilizam a estrutura de árvore possuem a vantagem de serem mais legíveis ao usuário, sem necessitar de um treinamento prévio.

Para ilustrar esse tipo de representação visual, são descritas as técnicas: (a) Grafos; (b) *Cone Trees*; (c) *Treemap*; (d) Empilhamento de dimensões e; (e) Gráfico de Mosaico.

a) Grafos

Grafos são objetos de estudo da teoria dos grafos. De forma geral, constitui-se de uma estrutura composta por vértices (ou nós) ligados por arestas (ou arcos). Como exemplos de aplicação de grafos, têm-se: na área biológica e química, a sua utilização para visualizar árvores genealógicas de espécies, mapas moleculares ou genéticos; visualizar a organização das páginas de um *web-site* e redes sociais; ilustrar problemas de busca da melhor rota, como o clássico problema do caixeiro viajante.

Outro exemplo de uso de grafos é a estrutura de dados denominada Árvore, definida como um grafo acíclico conectado. Sua estrutura é formada por nós que ramificam-se em vértices denominados filhos. Nós que não possuem filhos são definidos como nós-folhas. Este tipo de visualização é amplamente utilizado na representação de modelos de classificação, tornando sua interpretação mais intuitiva por meio da visualização da árvore (Urbanek, 2008; Wilkinson, 2008).

a.I) Tipo de tarefa

Os principais tipos de tarefa de grafos são visualizar e navegar pela estrutura de relacionamento formada pelos dados. Em (Wilkinson, 2008) e em (Nascimento e Ferreira, 2005) são brevemente abordados diferentes tipos de grafos como, por exemplo, árvore (grafo acíclicos) e rede (grafo cíclico).

a.II) Tipo de dado

Grafos representam tipicamente dados hierárquicos ou uma rede de dados, geralmente apresentados como uma estrutura de diretórios de um sistema de arquivos, taxonomias, estruturas organizacionais como de telecomunicações ou transporte, etc. (Wilkinson, 2008).

a.III) Dimensionalidade

Cada dimensão do dado pode ser atribuída a um nó, dependendo do objetivo da visualização. Dessa forma, um grafo pode representar dados com alta dimensionalidade (Oliveira e Levkowitz, 2003, p. 383).

a.IV) Volume

O tamanho do grafo é um fator crucial na visualização, pois pode comprometer a navegação entre os elementos e a distinção entre vértices e arestas. Assim, segundo Herman et al. (2000), a análise de um grafo é mais fácil para pequenas estruturas, logo, para um pequeno a médio volume de dados.

a.V) Posição

Se o grafo representa dados com hierarquias, então devem ser respeitados os níveis de hierarquia ao atribuir as dimensões dos dados (Keim, 2001).

b) Cone Trees

Esta técnica provê representação tridimensional para mapear dados hierárquicos formando uma estrutura de árvore com formato de cone translúcido, o que permite perceber todos os filhos dispostos na base do cone, bem como outros cones que podem estar atrás, como ilustra a Figura 4.32.a. A raiz da árvore encontra-se no vértice do cone cujos filhos são distribuídos na base, com espaçamentos iguais. Esta estrutura repete-se para cada nó da árvore que possua filhos.

Uma segunda alternativa é posicionar a árvore na horizontal, ilustrada na Figura 4.32.b, denominada *Cam Tree*. Recursos como animações, rotações, seleção, operações de ocultação ou exibição de nós, facilitam a exploração dos dados e o entendimento da hierarquia. Segundo Cockburn e McKenzie (2000), a animação empregada durante a rotação dos nós deve ser rápida, sendo necessária para que o usuário possa acompanhar a sequência de operações realizadas sobre a árvore.

(a) **(b)**

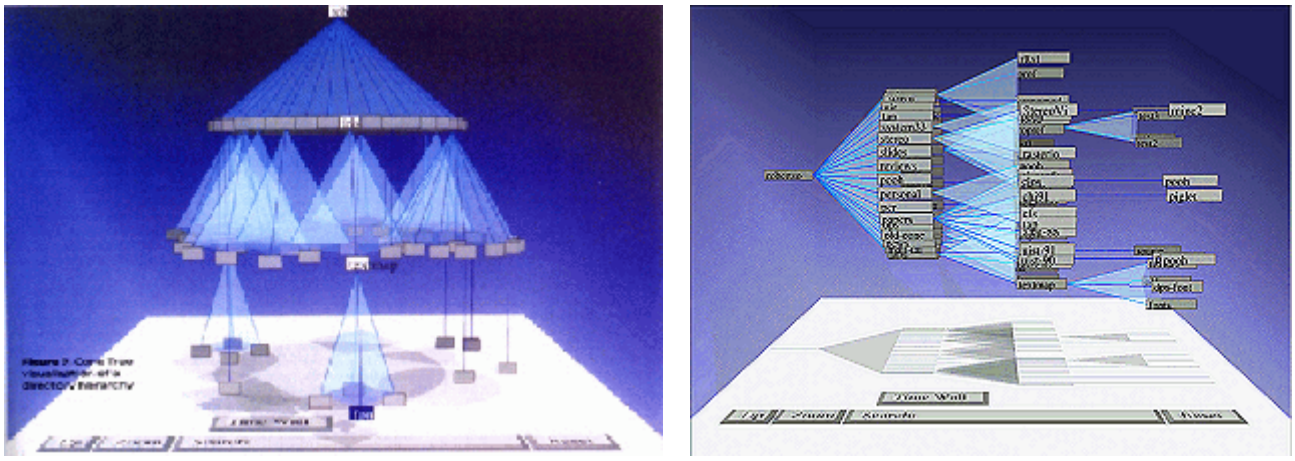

Figura 4.32: (a) Cone Tree (b) Cam Tree (Robertson et al., 1991)

b.I) Tipo de tarefa

Robertson et al. (1991) apontam que a principal vantagem dessa técnica é ser possível visualizar toda a estrutura da hierarquia de maneira mais simples, devido ao uso de rotações com efeitos de animação.

b.II) Tipo de dado

Essa técnica pode ser utilizada para manipular documentos de uma organização, como demonstra Robertson et al. (1991). Cockburn e McKenzie (2000) utilizaram esta técnica num experimento em que os dados representados foram organizados em uma estrutura hierárquica.

b.III) Dimensionalidade

Considerando que os nós representam as dimensões (ou atributos) dos dados, essa técnica pode representar dados com dimensionalidade alta. Segundo Robertson et al. (1991), a técnica *Cone Tree* pode chegar a representar 1000 nós. No trabalho apresentado por este autor, esta técnica foi utilizada para navegar pela estrutura de arquivos do sistema Unix, composto por 600 diretórios (nós) com um total de 10000 arquivos, que puderam ser visualizados inteiramente.

b.IV) Volume

Segundo Robertson et al. (1991), a técnica *Cone trees* possui um fator limitante de 30 pontos de ramificação ou 10 níveis de profundidade, de modo a não causar pontos de aglomeração na árvore, o que dificulta o entendimento do gráfico. Além disso, os autores acrescentam que a técnica demonstra-se mais eficiente quando a estrutura da hierarquia é desbalanceada, pois, dessa forma, o acompanhamento da sequência de rotações é mais fácil do que em árvores balanceadas, que sempre apresentam aparência uniforme. Entretanto, segundo Robertson et al. (1991), esta questão não é tão relevante uma vez que geralmente as hierarquias presentes nos dados são amplas, rasas (poucos níveis) e desbalanceadas, o que ameniza as limitações dessa técnica.

b.V) Posição

O posicionamento dos elementos de um determinado ramo da árvore não é relevante devido aos efeitos de animação providos pela técnica para navegar sobre uma estrutura de árvore previamente definida.

c) *Treemap*

Treemap representa a hierarquia presente nos dados em um espaço retangular, que é dividido em várias regiões, que por sua vez são divididas em outras, conforme os níveis da hierarquia a serem exibidos (Shneiderman, 2006). Todo o espaço do gráfico é aproveitado, permitindo observar de forma compacta a visualização completa de uma árvore de diretórios e arquivos (Johnson e Shneiderman, 1991), como é exemplificado na Figura 4.33.

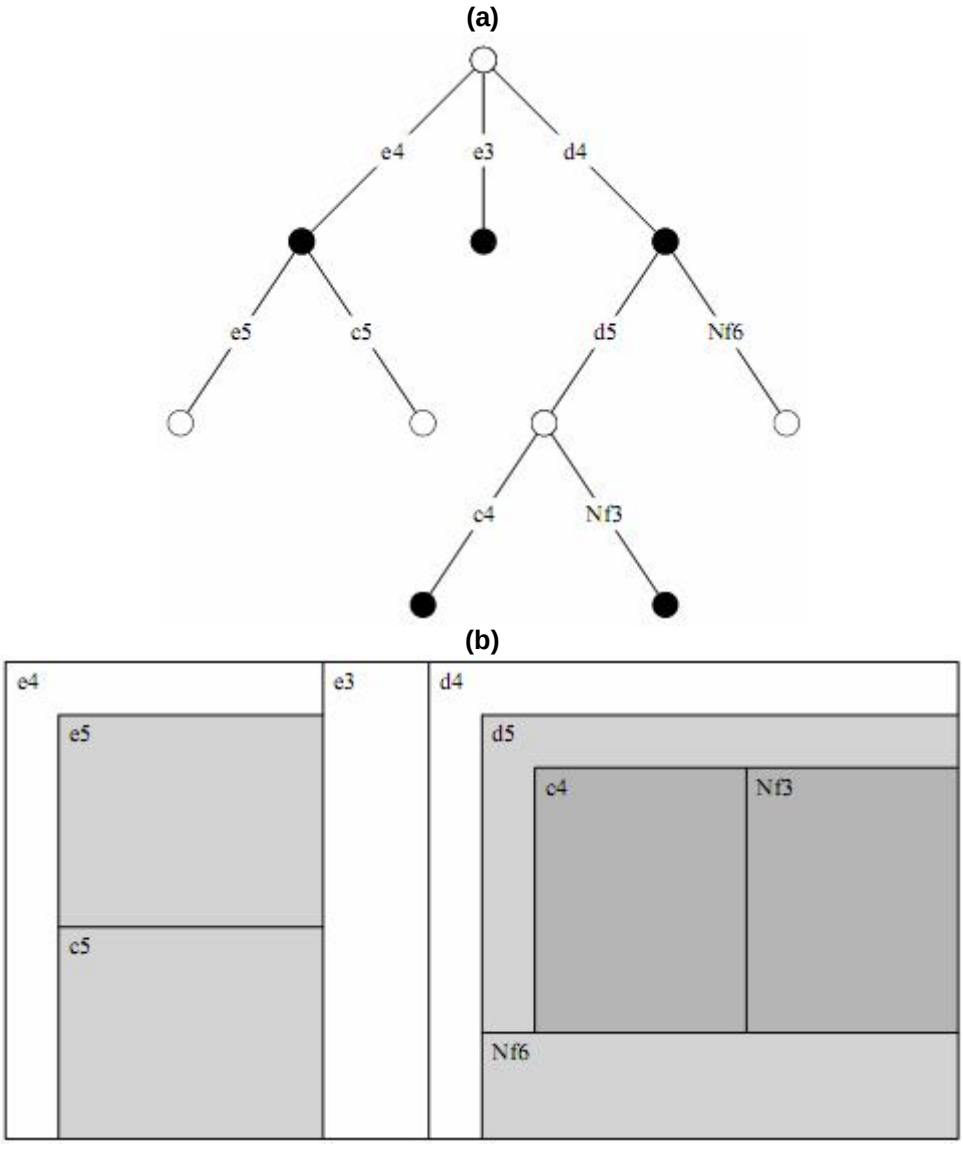

Figura 4.33: Princípio da técnica Treemap. (a) Representação tradicional de uma estrutura de uma árvore; (b) Treemap correspondente (Coulom, 2002)

c.I) Tipo de tarefa

Segundo Shneiderman (2006), a motivação original para desenvolver esta técnica era de visualizar o conteúdo de arquivos de um HD e verificar como cada arquivo estava armazenado e quanto de espaço estava sendo utilizado, para posteriormente poder detectar quais poderiam ser removidos. A visualização foi empregada para analisar dezenas de milhares de arquivos contidos em diretórios de 5 a 15 níveis.

Desde 1990, quando esta técnica foi concebida, até os dias atuais, sua utilização ampliou-se para outros domínios, como monitoramento de vendas (Shneiderman, 2006), visualização de moléculas (Baehrecke et al., 2004) e organização de álbuns de fotografia digital (Bederson, 2001). Nos três tipos de aplicações citadas, os dados possuem algum tipo de classificação e/ou hierarquia pré-estabelecida pelo domínio.

c.II) Tipo de dado

Assim como os grafos, *Treemaps* também representam a estrutura organizacional dos dados, seja hierarquia, taxonomia ou relação simples.

c.III) Dimensionalidade

A dimensionalidade, neste caso, é baseada nos níveis de hierarquia representados. Observando os trabalhos de Bederson (2001), (Baehrecke et al., 2004) e (Shneiderman, 2006), Treemaps podem exibir em torno de 2 a 20 regiões, que contêm os dados agrupados em uma mesma categoria.

c.IV) Volume

A exemplo dos trabalhos de Bederson (2001), (Baehrecke et al., 2004) e (Shneiderman, 2006), o volume de dados utilizado por esses autores é em volta de 100 a 3000 registros, compondo um volume médio de dados.

c.V) Posição

Bederson et al. (2002) apresentam vários algoritmos para posicionar os retângulos do Treemap. Esta técnica permite destacar elementos que possuem informações mais importantes, atribuindo uma região maior do espaço do gráfico. Dessa forma, o grau de relevância dos elementos pode ser verificado pelo tamanho do espaço que ocupam.

d) Gráfico de mosaicos (mosaic plots)

Tem esse nome devido ao fato de ser composto por um grupo de retângulos, cada qual representando um atributo categórico (ou qualitativo), cujo tamanho é proporcional ao valor correspondente ao atributo.

d.I) Tipo de tarefa

Fornece uma visão geral de dados qualitativos e possibilita a identificação de regras e padrões.

d.II) Tipo de dado

De acordo com Friendly (1998) e Hofmann (2008), o gráfico de mosaicos é adequado para representar dados qualitativos (ou categóricos).

d.III) Dimensionalidade

Baseado nos exemplos dados por Hofmann (2003), o diagrama de mosaicos pode representar dados com dimensionalidade de aproximadamente 10 atributos.

d.IV) Volume

Tomando os exemplos em (Hofmann, 2003), os dados representados possuem volume pequeno, pois são a respeito de eventos ou amostragem estatística.

d.V) Posição

O processo de construção desse gráfico é interferido pela ordem dos atributos, cujo arranjo pode enfatizar diferentes aspectos presentes nos dados (Hofmann, 2008). Um exemplo disso é dado por esse autor, que demonstra a construção de dois gráficos de mosaicos diferentes utilizando os mesmos dados, referentes às informações sobre os passageiros do famoso navio Titanic. A diferença desses gráficos está na ordem dos atributos *Classe* e *Sexo*, como indicam a Tabela 4.12 e a Tabela 4.13.

Tabela 4.12 : Dados sobre os passageiros do Titanic, organizados pela classe social (adaptado de (Hofmann, 2008))

Classe	Primeira		Segunda		Terceira		Tripulação	
Sexo	Feminino	Masculino	Feminino	Masculino	Feminino	Masculino	Feminino	Masculino
Sobreviveu:								
Sim	141	62	93	25	90	88	20	192
Não	4	118	13	154	106	422	3	670
Taxa de sobreviventes (em %)	97	34	88	14	46	17	87	22

Tabela 4.13: Dados sobre os passageiros do Titanic, organizados por sexo (adaptado de (Hofmann, 2008))

Sexo	Feminino				Masculino			
Classe	Primeira	Segunda	Terceira	Tripulação	Primeira	Segunda	Terceira	Tripulação
Sobreviveu:								
Sim	141	93	90	20	62	25	88	192
Não	4	13	106	3	118	154	422	670
Taxa de sobreviventes (em %)	97	88	46	87	34	14	17	22

As Figuras 4.33 e 4.34 ilustram, respectivamente, os dados das Tabela 4.12 e Tabela 4.13 (Hofmann, 2008). A taxa de sobreviventes está em destaque em ambos gráficos de mosaicos (do tipo tridimensional). Na Figura 4.34, pode-se observar que sobreviveu um número maior de mulheres. Na Figura 4.35 é enfatizada a taxa de sobreviventes entre as classes sociais, em que a taxa de mulheres sobreviventes aumenta com a classe e a taxa de homens sobreviventes é menor para a segunda classe.

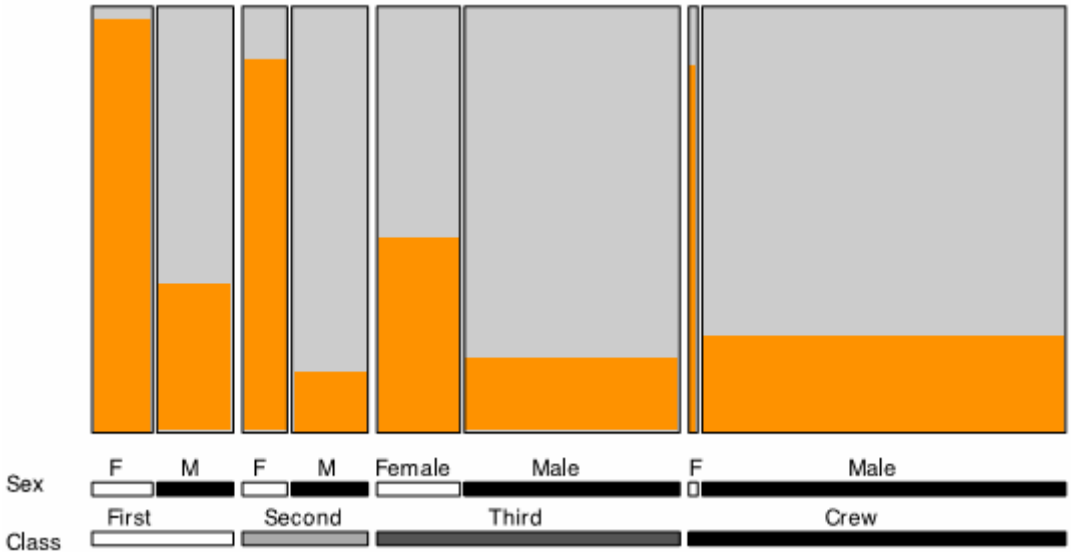

Figura 4.34: Gráficos de mosaico respectivos aos dados da Tabela 4.12(Hofmann, 2008; p.621)

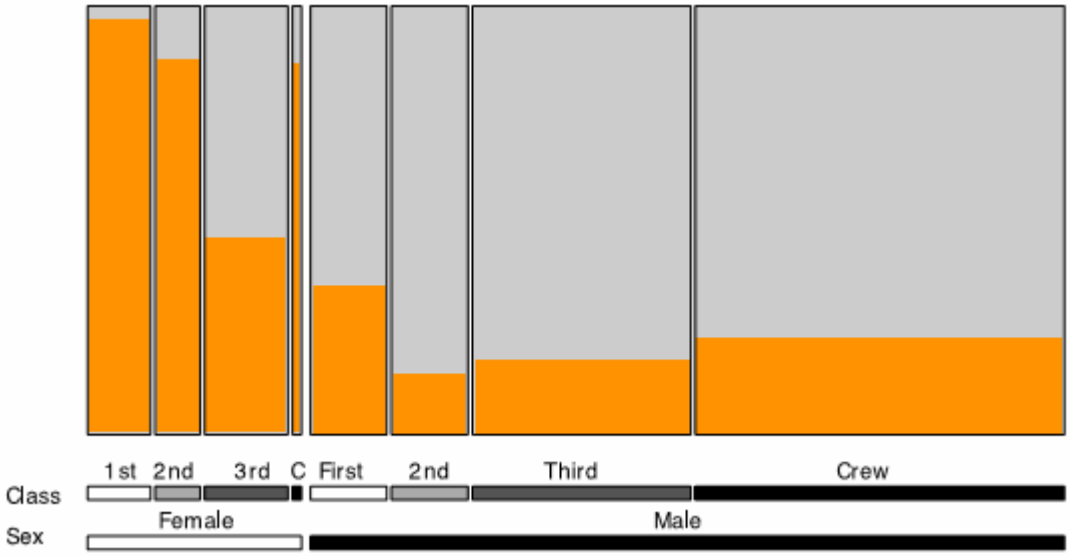

Figura 4.35: Gráficos de mosaico respectivos aos dados da Tabela 4.13(Hofmann, 2008; p.621)

e) Empilhamento de dimensões (*Dimensional Stacking*)

Exibe dados multidimensionais para um plano 2-D, em que as dimensões dos dados são encaixadas recursivamente umas nas outras, como ilustra a Figura 4.36. Cada dimensão do dado é dividida em uma escala de valores estipulada pelo usuário. Duas dimensões são atribuídas aos eixos *x* e *y*, marcados com essas escalas, formando uma grade. Dentro de cada grade é feito este mesmo procedimento para as próximas duas dimensões, que continua até todos os atributos serem mapeados.

Segundo LeBlanc et al. (1990), a técnica Empilhamento de Dimensões foi criada para a visualização de dados com mais de três dimensões, superando os gráficos estatísticos (em que os

dados têm que passar por um processo de transformação) e sendo uma alternativa às técnicas iconográficas.

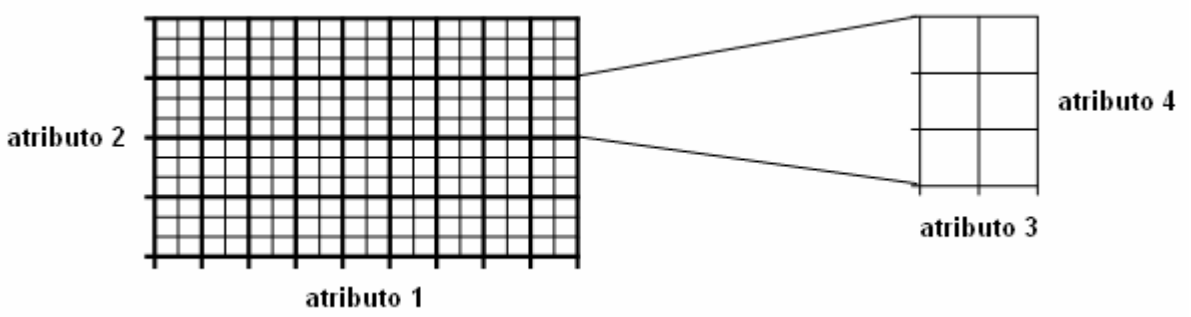

Figura 4.36: Exemplo de particionamento das dimensões do dado na técnica de Empilhamento de dimensões (adaptado de (Keim, 1997))

e.I) Tipo de tarefa

A técnica empilhamento de dimensões é utilizada para encontrar agrupamentos, padrões, *outliers* ou regras presentes nos dados (Hoffman, 1999). Como exemplo de aplicação, Taylor et al. (2006) utilizaram a técnica de empilhamento de dimensões para visualizar o comportamento dos parâmetros de um modelo neural e facilitar a inferência sobre este. A base de dados avaliada possui oito dimensões (atributos): *KCa*; *Na*; *Kd*; *CaT*; *h*; *CaS*; *leak*; e *A*.

De acordo com Taylor et al. (2006), cada atributo contém valores que podem variar entre seis escalas diferentes (que no domínio da aplicação são denominados como valores de condutância[5] dos elementos). A Figura 4.37 ilustra diferentes combinações obtidas de acordo com a variação das ordens de empilhamento dos atributos e seus valores para as dimensões do gráfico. Durante este mapeamento, cada elemento foi designado para uma cor diferente, o que permitiu a visulização de padrões de comportamento pelo gráfico gerado. Assim, Taylor et al. (2006) puderam visualizar a relação entre os oito atributos e seus respectivos valores de condutâncias e analisar o comportamento do modelo neural.

A Figura 4.38 ilustra as etapas de construção de uma dessas combinações. Na parte da Figura 4.38 indicada pela letra A, são visualizados dois atributos (*KCa* e *Na*) em relações aos demais (*Kd, CaT, h, CaS, leak, A*) cujos valores de condutância foram colocados em zero, formando uma grade 6x6. Na parte B, são visualizados os próximos dois atributos, *Kd* e *CaT*, em relação aos dois primeiros, formando uma grade 36x36, permanecendo o restante dos

[5] Do inglês *conductance*: na Eletrônica, refere-se à medida da facilidade com que um dado condutor permite uma corrente elétrica circular por ele, sendo o inverso da resistência. (fonte: Michaelis – Moderno Dicionário da Língua Portuguesa)

atributos (*h, CaS, leak, A*) com valores de condutância iguais a zero. As partes C e D da Figura 4.38 mostram a repetição deste procedimento para os demais atributos de modo que, ao final, todos sejam mapeados no gráfico.

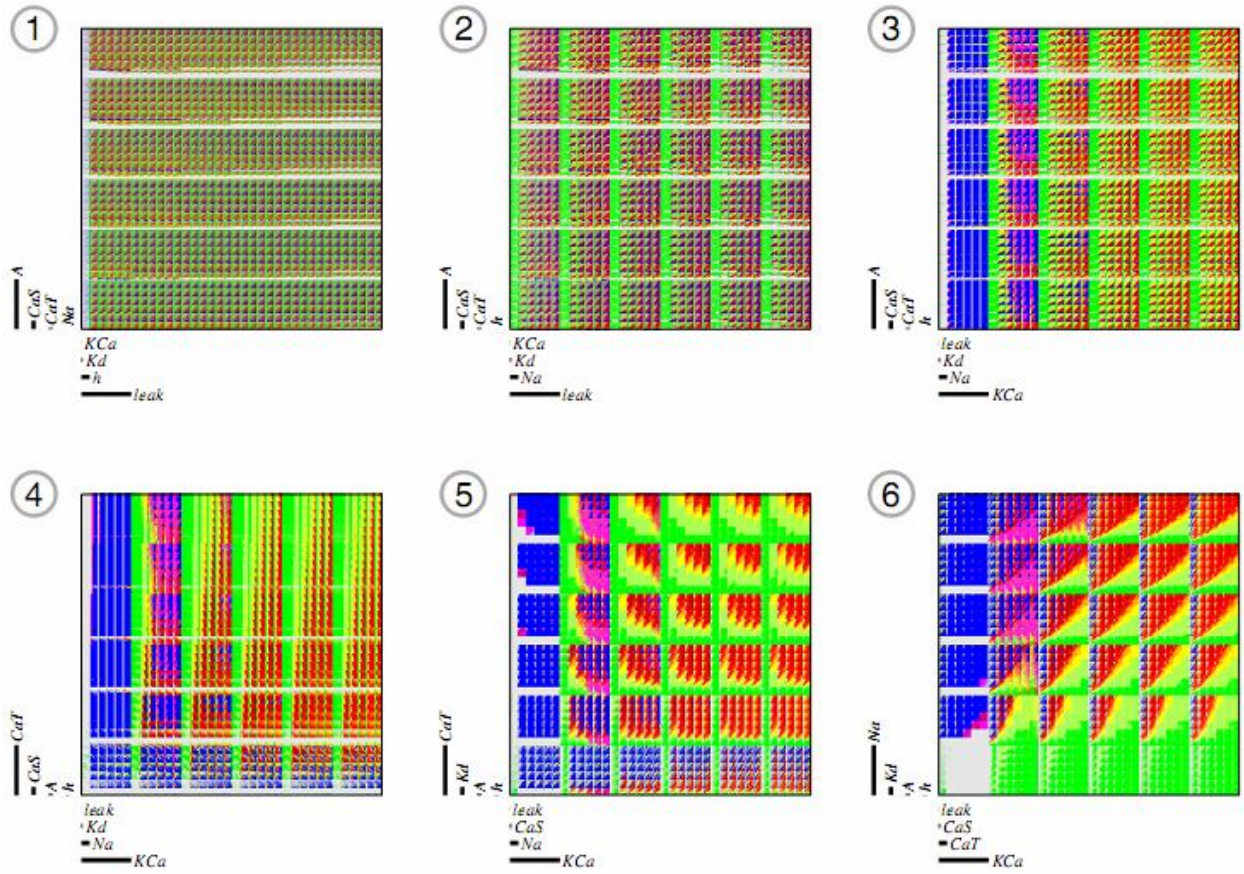

Figura 4.37: Visualização do comportamento dos atributos do exemplo de aplicação (Taylor et al., 2006)

O exemplo do trabalho de Taylor et al. (2006), ilustra a importância da interação entre o analista e a visualização durante a exploração dos dados. A aplicação da técnica Empilhamento de Dimensões serviu para analisar o comportamento de um modelo, expresso por meio de padrões formados conforme as diferentes ordenações dos atributos.

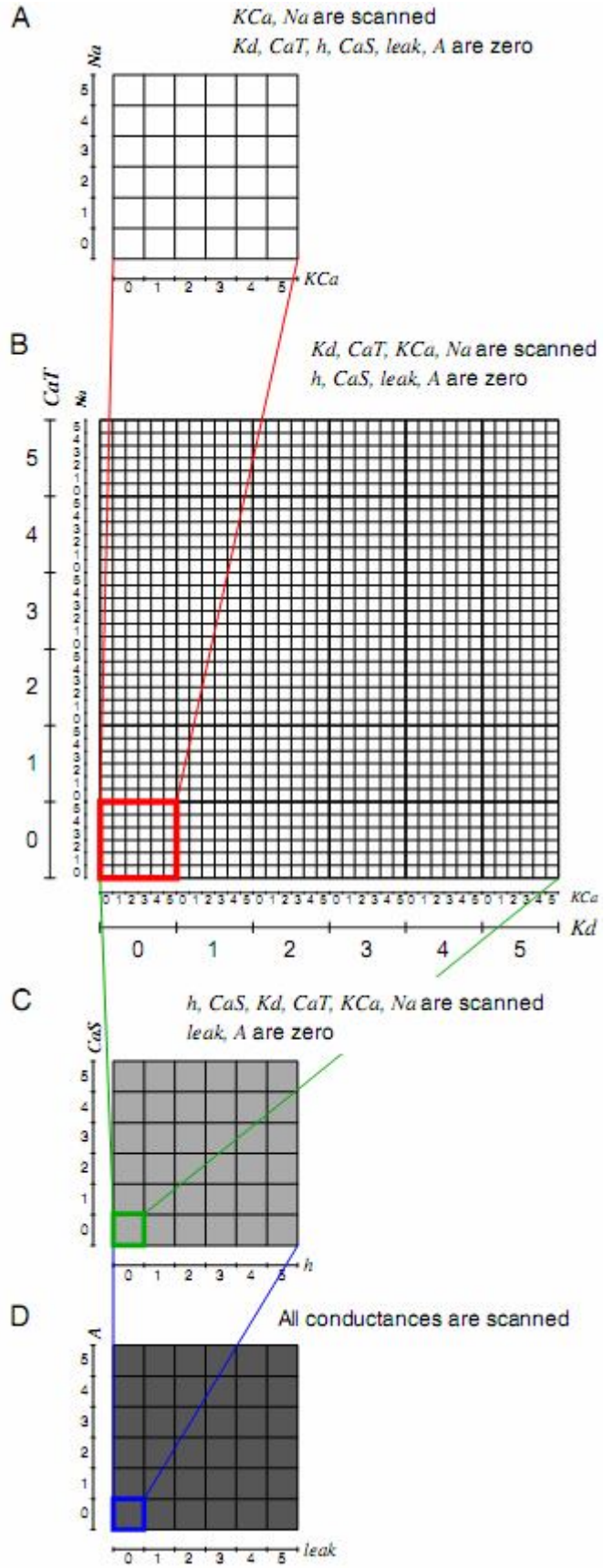

Figura 4.38: Exemplo de utilização da técnica empilhamento de dimensões (Taylor et al., 2006)

e.II) Tipo de dado

Pelos exemplos apresentados em (LeBlanc et al., 1990), (Taylor et al., 2006) e Keim (2002), pode-se verificar que os dados representados devem ser quantitativos discretos ou contínuos, cujos valores são distribuídos recursivamente para os eixos do gráfico

e.III) Dimensionalidade

Segundo Hoffman (1999, p.29), a divisão da escala (quantidade de células da grade) não deve ultrapassar mais que cinco categorias e é desejável que não mais de nove atributos sejam utilizados para não dificultar a visualização. Cada registro da base de dados é atribuído a uma célula da grade que corresponde ao valor da escala e pode ser distinto por uma cor.

e.IV) Volume

Segundo Ward et al. (1994), um volume pequeno de dados com alta dimensionalidade pode causar visualizações vazias; por outro lado, grande quantidade de dados com alta dimensionalidade também pode causar consumo de tempo elevado para a análise dos dados representados. Portanto, o conjunto de dados ideal para essa técnica de visualização deve ter uma dimensionalidade média (cinco a nove atributos) e um volume médio de dados.

e.V) Posição

Peng et al. (2004) ressaltam que a ordem da distribuição das dimensões causa grande impacto na visualização. Keim (2002) observa ainda que é viável distribuir primeiro os atributos de maior importância para as extremidades das grades, colocando as de menor relevância no interior. Como visto anteriormente na descrição do exemplo do trabalho de Taylor et al. (2006), a posição dos atributos é um parâmetro que influencia fortemente na análise dos dados.

4.5. Processo de Exploração Visual Combinando Diferentes Técnicas

Durante a exploração visual, é interessante utilizar mais de uma técnica de visualização visando a obtenção de diferentes perspectivas para os mesmos dados. São apresentados nesta seção trabalhos relacionados à integração de diferentes técnicas de visualização.

Wills (2008) faz uma abordagem sobre *linked views*, técnica para comunicar duas ou

mais visualizações sobre um mesmo conjunto de dados de modo que uma modificação realizada em uma das visualizações seja refletida nas demais. Um exemplo é o destaque de determinado conjunto de dados exibido simultaneamente em todos os gráficos gerados. A ligação entre as visualizações é um meio que o analista possui para explorar os dados sob diferentes pontos de vista e interagir dinamicamente sobre eles.

Outra abordagem semelhante à de *linked views*, denominada *Coordinated and Multiple Views* (CMV), refere-se a uma técnica que permite ao usuário explorar seus dados por meio de interações com diferentes representações visuais que proporcionam melhor entendimento ao apresentá-los sob diferentes perspectivas com o uso de múltiplas técnicas de visualização (Roberts, 2007).

A diferença entre os conceitos apresentados em trabalhos sobre *linked views* e CMV está, basicamente, relacionada à contemplação de tipos de técnicas de visualização utilizados na representação dos dados.

Para ambas abordagens, a comunicação entre as diferentes visualizações sobre um mesmo conjunto de dados deve ser viabilizada por um software com capacidade de compartilhar e controlar os dados nessas visualizações. Wilhelm (2008) e Andrienko e Andrienko (2007) citam alguns softwares que oferecem esta funcionalidade. Alguns exemplos desse tipo de software são descritos no Capítulo 5.

Além disso, existem trabalhos com o intuito de aproveitar as vantagens de cada técnica combinando-as em uma única, como exemplos apresentados na Seção 4.5.2.

4.5.1. *Linked Views vs. Coordinated and Multiple Views*

Na literatura encontram-se os termos *linked views* e *coordinated and multiple views* que, quando confrontados, apresentam mais semelhanças do que diferenças. Trabalhos relacionados a ambos os conceitos convergem no que diz respeito à utilização de diferentes tipos de técnicas de visualização na representação de dados, sobre os quais a interação do usuário realizada em uma visão deve ser refletida para as outras.

Estes paradigmas são concernentes à construção de software de visualização de dados quando um dos requisitos do projeto é a integração entre as diferentes técnicas de visualização implementadas no software, com a finalidade de proporcionar flexibilidade durante a exploração de dados.

Wilhelm (2008) discute vários aspectos da abordagem *linked views*, como a representação de diferentes estruturas de dados e estratégias ou funcionalidades oferecidas pelo

software para a exploração de dados.

Wills (2008) apresenta exemplos de integração entre duas ou mais técnicas de visualização da classe de gráficos 1D-3D, técnicas geométricas e técnicas hierárquicas, com a finalidade de demonstrar técnicas e decisões de projeto de implementação de um software a ser construído utilizando o paradigma *linked views*.

Roberts (2007) realizou um estudo sobre CMV abordando aspectos inerentes à exploração visual de dados por meio de sistemas que implementam este paradigma, como a preparação dos dados, geração e controle de múltiplas visualizações, técnicas de interação e manipulação, exemplos de ferramentas e a interface com o usuário final.

Baldonado et al. (2000) estabelecem diretrizes para a construção de sistemas que utilizem múltiplas visualizações abordando aspectos em relação a decisões de projeto e seus respectivos impactos.

A Tabela 4.14 reúne a descrição dos conceitos apresentados anteriormente, de acordo com os autores referenciados na terceira coluna desta tabela.

Tabela 4.14: Descrição dos conceitos linked view e coordinated and multiple views

Conceito	**Descrição**	**Referência**
Linked views	Utilização de dois ou mais gráficos que compartilham e intercambiam informações	(Wilhelm, 2008)
	Integração entre técnicas de visualização que compartilham o mesmo conjunto de dados para representar diferentes seus aspectos sob o ponto de vista de cada tipo de técnica utilizada.	(Wills, 2008)
Coordinated and multiple views	Tem como foco a busca de informações nos dados por meio de um conjunto integrado de técnicas de visualização	(Roberts, 2007)
Multiple view system	Sistema que utiliza duas ou mais técnicas de visualização distintas para apoiar a exploração de um mesmo conjunto de dados, fornecendo diferentes informações sobre estes	(Baldonado et al., 2000)

Observando-se as definições de cada terminologia descritas na Tabela 4.14, pode-se afirmar que todas referem-se a combinações integradas de diferentes técnicas de visualização, sendo cada tipo de técnica responsável por apresentar um aspecto específico dos dados, proporcionando, desse modo, uma visão multifacetada destes. As técnicas de visualização geralmente utilizadas são da classe de gráficos 1D a 3D, técnicas geométricas, mapas e técnicas hierárquicas.

4.5.2. Combinação de técnicas

Fanea et al. (2005) fazem a combinação das coordenadas paralelas com S*tar Glyphs*, denominada *Parallel Glyphs*, em que as linhas poligonais das coordenadas paralelas conectam cada ícone de *Star Glyph*, como ilustra a Figura 4.39. Segundo estes autores, esta técnica reduz o problema da sobreposição de dados das coordenadas paralelas, em que as linhas poligonais, ao invés de interceptar eixos verticais fixos, intercepta segmentos individuais, que originam-se a partir de um ponto central, assim como funciona na técnica *Star Glyph*. Desse modo, as coordenadas paralelas podem ser vistas a partir de uma visão em 3D. Na Figura 4.40 é ilustrado um exemplo de coordenadas paralelas convencional, em que a linha poligonal A é parcialmente encoberta por outras. Quando é aplicada a nova técnica, como ilustra a Figura 4.41, a linha A pode ser visualizada individualmente e, além disso, pode-se observar que outra linha poligonal C estava encoberta pela sobreposição da visão em 2D.

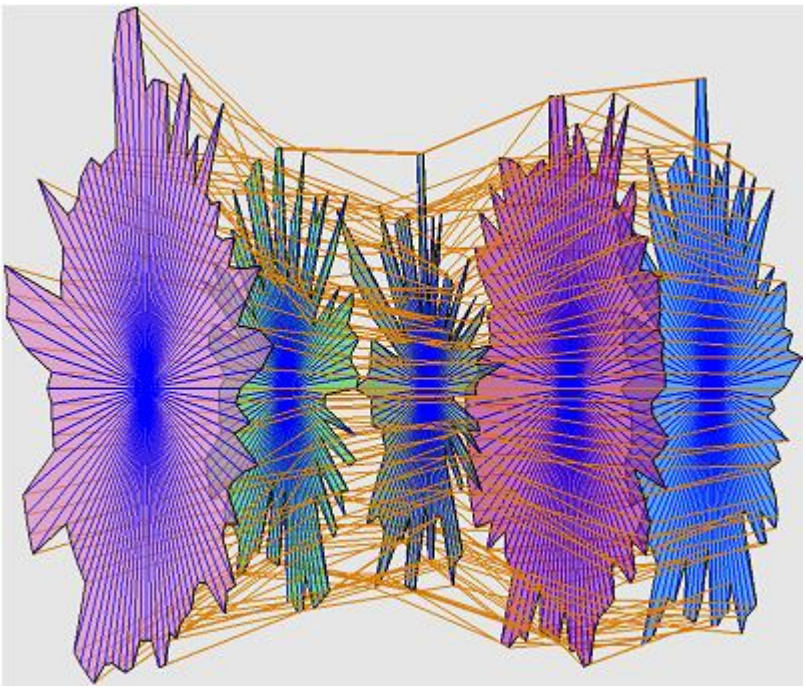

Figura 4.39 : Combinação das técnicas Coordenadas Paralelas e Star glyphs (Fanea et al.,2005)

Figura 4.40: Coordenadas paralelas tradicionais (Fanea et al.,2005)

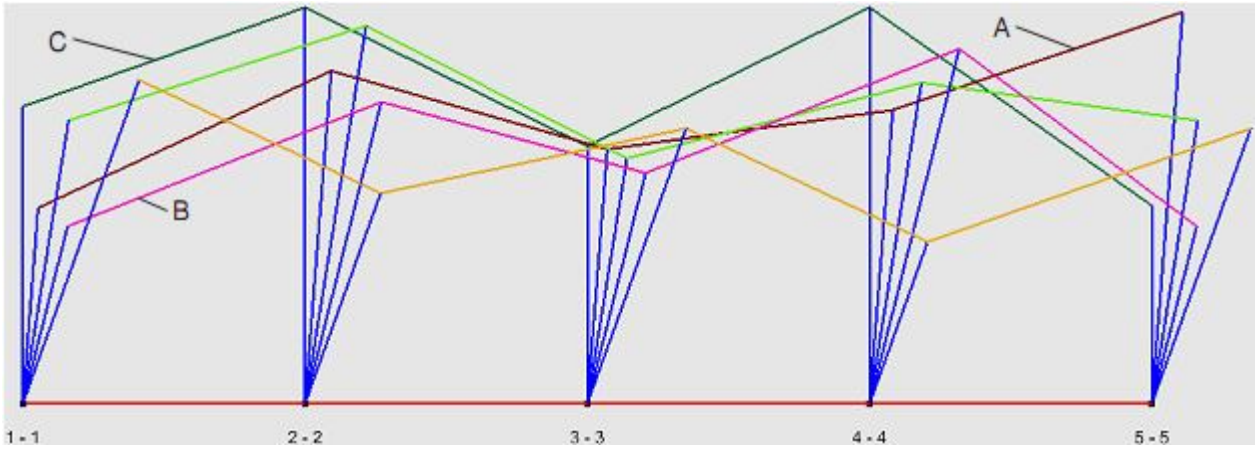

Figura 4.41: Combinação de Coordenadas paralelas com Star glyphs (Fanea et al.,2005)

Keim et al. (2004) apresentam uma nova técnica, chamada *Circle View*, que combina a visualização hierárquica, como *Treemaps*, e técnicas de formatos circulares, como gráfico de pizza e *Circle segments*. O objetivo desta nova técnica, ilustrada pela Figura 4.42, é representar dados contínuos com características temporais, típicos de sistemas de tempo real. A ideia básica é "representar as dimensões dos dados por segmentos de um círculo, em que cada segmento é dividido em sub-segmentos indicando as distribuições e mudanças ao longo do tempo" (Keim et al., 2004; p. 179).

Figura 4.42: Técnica Circle Views (Keim et al., 2004)

4.6. Considerações finais

Neste capítulo foram apresentadas algumas técnicas de visualização com o objetivo de ilustrar a análise dos parâmetros sobre os quais deve-se estar atento na escolha dessas técnicas. Os parâmetros analisados foram: tipo de dado, tipo de tarefa, volume e dimensionalidade dos dados e posição dos atributos.

O tipo dos dados é o fator primordial para determinar o tipo de técnica. Dados qualitativos possuem um número menor de opções de técnicas quando comparados aos dados quantitativos, que podem ser representados por todas as técnicas apresentadas neste trabalho. Algumas, como a técnica orientada a pixel *query-independent* e figuras de arestas, são apropriadas para dados que apresentam características como, por exemplo, a marcação de tempo e/ou espaço e dados físicos. Dados que seguem naturalmente algum tipo de hierarquia ou relacionamento são melhores representados por técnicas baseadas em grafos ou hierárquicas.

A escolha da técnica também pode ser influenciada pelo objetivo do usuário na utilização da visualização. Pode ser que o usuário tenha interesse em saber a correlação entre determinados atributos ou ter uma visão geral de todos os dados para, a partir de então, filtrar um subconjunto de interesse e fazer uma exploração mais profunda. Ou então, o usuário deseja identificar regras, padrões de comportamento, agrupamentos ou anomalias presentes nos dados.

Outro fator a ser observado é a quantidade de dimensões e o volume de dados que a técnica é capaz de comportar. Técnicas para representar dados com uma a três dimensões comportam um número reduzido de dados, pois geralmente são utilizadas para visualizar um conjunto de dados que são resultado de um cálculo estatístico. Todas as demais técnicas apresentadas são aptas a representar dados multidimensionais, com mais de três atributos.

Técnicas iconográficas, em geral, representam um número menor de atributos do que técnicas geométricas, pois estão limitadas pelas propriedades que os ícones dispõem para representar os atributos dos dados. Além disso, a quantidade de dados representada é limitada pela capacidade do espaço disponível da visualização para exibir todos os ícones, que podem sofrer diminuição do tamanho, o que pode afetar a identificação de suas propriedades. Consequentemente, técnicas iconográficas comportam um volume pequeno de dados.

Com técnicas geométricas é possível representar um volume médio de dados, de modo a não formar "borrões" formados pela sobreposição dos registros representados. As técnicas orientadas a pixel não destacam-se pela quantidade de atributos que podem representar – uma quantidade média a alta – mas sim, pelo grande volume de dados. Já as técnicas baseadas em grafos não focam a dimensionalidade dos dados, sendo voltadas para representar o relacionamento entre seus atributos. Semelhante às técnicas iconográficas, as técnicas baseadas em grafos podem ter o espaço de exibição como um fator limitante para a representação de todos os dados, sendo cabível a utilização de técnicas de interação, como a expansão de nós ou *zoom*, caso o volume de dados seja muito grande.

A posição dos atributos na visualização também é um aspecto a ser levado em consideração, pois a ordem de atribuição das dimensões pode facilitar ou dificultar a identificação de padrões ou regras presentes nos dados. Isto é válido para as classes de técnicas iconográficas, geométricas e orientadas a pixel. Por isso, é recomendável posicionar próximos os atributos sobre os quais deseja-se averiguar algum tipo de regra, padrão ou relacionamento.

Keim e Kriegel (1996) observam que a avaliação de técnicas de visualização para indicar se são ou não adequadas na representação dos dados, é um tanto quanto subjetiva, pois está ligada ao fator humano e à sua percepção e facilidade de interpretação sobre as características de cada técnica.

5. Triangulação: utilização de ferramentas computacionais para a visualização de dados

5.1. Considerações iniciais

Segundo Wainer (2007), triangulação é um método da pesquisa qualitativa observacional que consiste em “buscar pelo menos duas formas/fontes para cada dado e análise da pesquisa. Pode-se usar mais de uma técnica de coleta de dados, por exemplo, análise de documentos e entrevistas, ou observação e questionários, ou pode-se usar mais de um pesquisador observando o ambiente”.

A triangulação é utilizada neste trabalho como um complemento da Teoria Fundamentada em Dados, diversificando a fase de coleta de dados para que não seja baseada somente na literatura, ampliando assim a amostragem teórica, conforme sugerido por Wainer (2007). Desse modo, a forma de triangulação utilizada é a análise e observação sobre o resultado gerado por ferramentas de visualização existentes, que implementam as técnicas de visualização descritas no Capítulo 4.

Neste capítulo são apresentados exemplos de aplicação de algumas técnicas de visualização seguindo as observações realizadas durante a identificação dos parâmetros definidos no capítulo anterior. Para isto, foram pesquisadas algumas ferramentas que podem ser utilizadas para análise de bases de dados, descritas na Seção 5.2. Em seguida, são apresentadas as bases de dados utilizadas e respectivas análises dos resultados da visualização para cada uma delas. Com isso, pretende-se confirmar a importância de considerar os parâmetros levantados para a escolha de técnicas de visualização apropriadas aos dados sob análise.

A avaliação de ferramentas não faz parte dos objetivos deste trabalho, sendo este assunto

abordado por trabalhos como os de Chi (2000), que descreve uma taxonomia de técnicas de visualização examinando o domínio compatível com as técnicas de visualização, e de Valiati (2008), que apresenta uma avaliação das técnicas de visualização de acordo com as tarefas que o usuário pode realizar sobre a visualização, durante a análise de dados. Além disso, este capítulo não propõe-se a apresentar um determinado problema e tentar resolvê-lo utilizando técnicas de visualização. O intuito neste caso é demonstrar a análise das características presentes nos dados e aplicar as técnicas de visualização que são possíveis conforme os parâmetros estabelecidos.

Convém observar que não foi possível a análise de todas as técnicas apresentadas durante o capítulo anterior, em razão de dificuldades de acesso a ferramentas que implementam determinadas técnicas, ou ao tipo de dados analisados não ser próprio para uma técnica específica.

5.2. Softwares interativos para análise visual

Para realizar a análise visual de dados, é necessária a utilização de ferramentas que implementem técnicas de visualização. Na Tabela 5.1 estão especificadas alguns exemplos dessas ferramentas, suas respectivas fontes e quais técnicas implementam. Essas ferramentas foram desenvolvidas para a exploração e análise de dados multidimensionais. Foi dada preferência àquelas disponíveis gratuitamente, mas em <http://www.kdnuggets.com/software/visualization.html> (acessado em 16 de dezembro de 2009) há uma lista contendo outras ferramentas comerciais e *on-line*, dentre tantas que também podem ser encontradas na Internet.

Neto (2008) classifica os softwares descritos na Tabela 5.1 como Sistemas Interativos, em que o usuário pode ver alterações dos parâmetros refletidas na visualização de forma interativa, não exigindo conhecimento de programação. Ferramentas como a Linguagem R, utilizada por Rabelo (2007), dentre outras como Maple, MatLab e Graphviz são classificadas por Neto (2008) como Sistemas Interpretativos, pois a visualização dos dados é feita por meio da interpretação de *scripts* provenientes de arquivos textos ou linhas de comandos disponíveis no software. Esta classe de ferramentas não é tratada neste trabalho. A seguir são descritas, de forma mais detalhada, as ferramentas relacionadas na Tabela 5.1.

Tabela 5.1: Ferramentas disponíveis e técnicas de visualização que implementam

Ferramenta	**Disponível em**	**Técnicas implementadas**
CASSAT	http://stats.math.uni-augsburg.de/CASSATT/ multiplataforma (aplicativo Java)	Diagrama de caixas Coordenadas paralelas Gráfico de dispersão
GAUGUIN	http://stats.math.uni-augsburg.de/software/Gauguin/gauguin.html para Windows, Linux e Mac OS	*Star Glyphs* Matriz de dispersão Histograma Diagrama de barras
HCIL	http://www.cs.umd.edu/hcil/research/visualization.shtml	Grafos *Treemaps*
KLIMT	http://stats.math.uni-augsburg.de/KLIMT/intro.html para Windows, Linux e Mac OS	Árvores *Treemaps* Histograma Gráfico de barras Gráfico de dispersão Gráfico de caixa
Many Eyes	http://manyeyes.alphaworks.ibm.com/manyeyes/ ferramenta *on-line*	Diagramas 1-D a 3-D Diagrama de dispersão Grafos *Treemaps* Gráficos para análise de textos Mapas
Mondrian	http://stats.math.uni-augsburg.de/Mondrian/ para Windows, Linux e Mac OS	Histograma Diagrama de mosaicos Diagrama de barras Diagrama de dispersão Diagrama de caixa Mapas Coordenadas paralelas
Parvis	http://home.subnet.at/flo/mv/parvis/index.html multiplataforma (aplicativo Java)	Coordenadas paralelas
XmdvTool	http://davis.wpi.edu/~xmdv/ para Windows	Gráfico de dispersão *Star Glyphs* Coordenadas paralelas Empilhamento de dimensões Orientada a pixel

CASSAT, GAUGUIN, KLIMT e Mondrian são softwares desenvolvidos pelo grupo de pesquisa do departamento de estatística computacional e análise de dados da Universidade de Augsburgo, Alemanha. CASSAT implementa a técnica de coordenadas paralelas e suas variações. GAUGUIN implementa técnicas para analisar dados categóricos e contínuos como *star glyphs*, diagrama de dispersão, diagrama de barras e histograma. KLIMT foi desenvolvida para visualizar o relacionamento de dados hierárquicos, dispondo-os na estrutura de árvore, que

pode ser manipulada interativamente. Mondrian é um software de visualização de dados para propósitos estatísticos. Implementa técnicas como diagrama de mosaicos, diagrama de barras, coordenadas paralelas, diagrama de caixas, diagrama de dispersão, matriz de dispersão, histograma, mapas e outros gráficos para análise de inconsistências nos dados (dados nulos). Além disso, coordenadas paralelas e diagrama de caixas podem ser combinados para exploração dos dados (Theus, 2002).

HCIL (*Human-Computer Interaction Lab*) é um grupo de pesquisa sobre a área Interação Humano-Computador, da Universidade de Maryland, EUA, que possui a Visualização de Dados como um dos campos de estudo e desenvolve projetos que envolvem principalmente técnicas de visualização baseadas em grafos e *Treemaps*. No endereço eletrônico referente ao HCIL, citado na Tabela 5.1, são encontrados diversos aplicativos desenvolvidos pela equipe.

Many Eyes é uma ferramenta *on-line*, que assim como a Mondrian, implementa técnicas de visualização que auxiliam a análise estatística dos dados.

Parvis, por sua vez, é dedicada à técnica de Coordenadas paralelas e, por fim, a ferramenta XmdvTool, referenciada por Keim (2001), implementa as técnicas: gráfico de dispersão, s*tar glyphs*, coordenadas paralelas, empilhamento de dimensões e orientada a pixel.

5.3. Aplicação de técnicas de visualização sobre bases de dados

Nesta seção são analisadas bases de dados para exemplificar o emprego das técnicas de visualização segundo os parâmetros levantados: tipo de dado, tipo de tarefa, volume, dimensionalidade e posicionamento. A princípio, são verificados os parâmetros que podem ser detectados diretamente, como tipo de dado, volume e dimensionalidade. O tipo de tarefa depende do objetivo do analista e da ferramenta utilizada. O posicionamento deve ser observado de acordo com a técnica de visualização.

As bases utilizadas referem-se a: dados sobre automóveis; informações sobre passageiros do Titanic; inscritos no vestibular de 2009 da Universidade Estadual de Maringá; dados sobre a investigação de depressão pós-parto. Para cada base de dados é feita:

(a) caracterização dos dados: inicialmente a base de dados é analisada por meio dos parâmetros: dimensionalidade, volume e tipo de dado. Os parâmetros referentes à tarefa e posicionamento dos atributos são analisados após a aplicação das ferramentas de visualização;

(b) aplicação de técnicas de visualização: são selecionadas e aplicadas as possíveis técnicas mais adequadas às características dos dados diagnosticadas em (a);

(c) análise dos resultados: é feita uma discussão geral sobre os resultados obtidos das ferramentas de visualização utilizadas, incluindo a análise dos tipos de tarefas realizadas, considerando as descritas na Seção 4.3.2, e também a relevância do posicionamento dos atributos no gráfico gerado.

5.3.1. Base de dados de automóveis

Esta base de dados está disponível em <http://davis.wpi.edu/xmdv/datasets.html>, e é utilizada para ilustrar exemplos de técnicas de visualização implementadas pela ferramenta Xmdv.

a) Características dos dados

A base de dados descrita na Tabela 5.2 é composta principalmente por valores quantitativos e apresenta volume pequeno e média dimensionalidade. Por isto, podem ser empregadas, por exemplo, técnicas iconográficas para comparar os atributos e técnicas geométricas para analisar o comportamento dos mesmos.

Tabela 5.2: Base de dados de automóveis europeus, americanos e japoneses, dos anos 70 a 90

Parâmetros	**Valor**	**Características**
Número de dimensões	7	dimensionalidade média
Quantidade de registros	392	volume pequeno
Descrição dos atributos	MPG (milhas por galão)	quantitativo contínuo
	Cilindros (*Cylinders*)	quantitativo discreto
	Cavalos de potência (*Horsepower*)	quantitativo contínuo
	Peso (*Weight*)	quantitativo contínuo
	Aceleração (*Acceleration*)	quantitativo contínuo
	Ano (*Year*)	quantitativo discreto
	Origem (EUA, Japão, Europa)	qualitativo nominal

b) Aplicação de técnicas de visualização

No endereço eletrônico <http://davis.wpi.edu/xmdv/cs_cars.html> (acessado em 27 de dezembro de 2009) há um exemplo de aplicação das técnicas de visualização implementadas pela

ferramenta Xmdv, o qual foi reproduzido neste trabalho, utilizando a mesma ferramenta, cujos resultados são ilustrados pelas Figuras 5.1, 5.2, 5.3, 5.4 e 5.5.

A Figura 5.1 mostra a seleção dos registros de carros com maior valor de milhas por galão (MPG), utilizando a técnica Coordenadas Paralelas. A seleção dos registros pode ser observada pela área sombreada atrás das linhas poligonais, que ganham destaque por meio de uma cor mais forte. Nota-se que a maioria dos selecionados possui quatro cilindros e baixo peso. Pode-se observar também que há dois registros com valor de cilindro maiores (5 e 6 cilindros) que também apresentam bom aproveitamento de combustível, pois estão entre os registros do grupo selecionado.

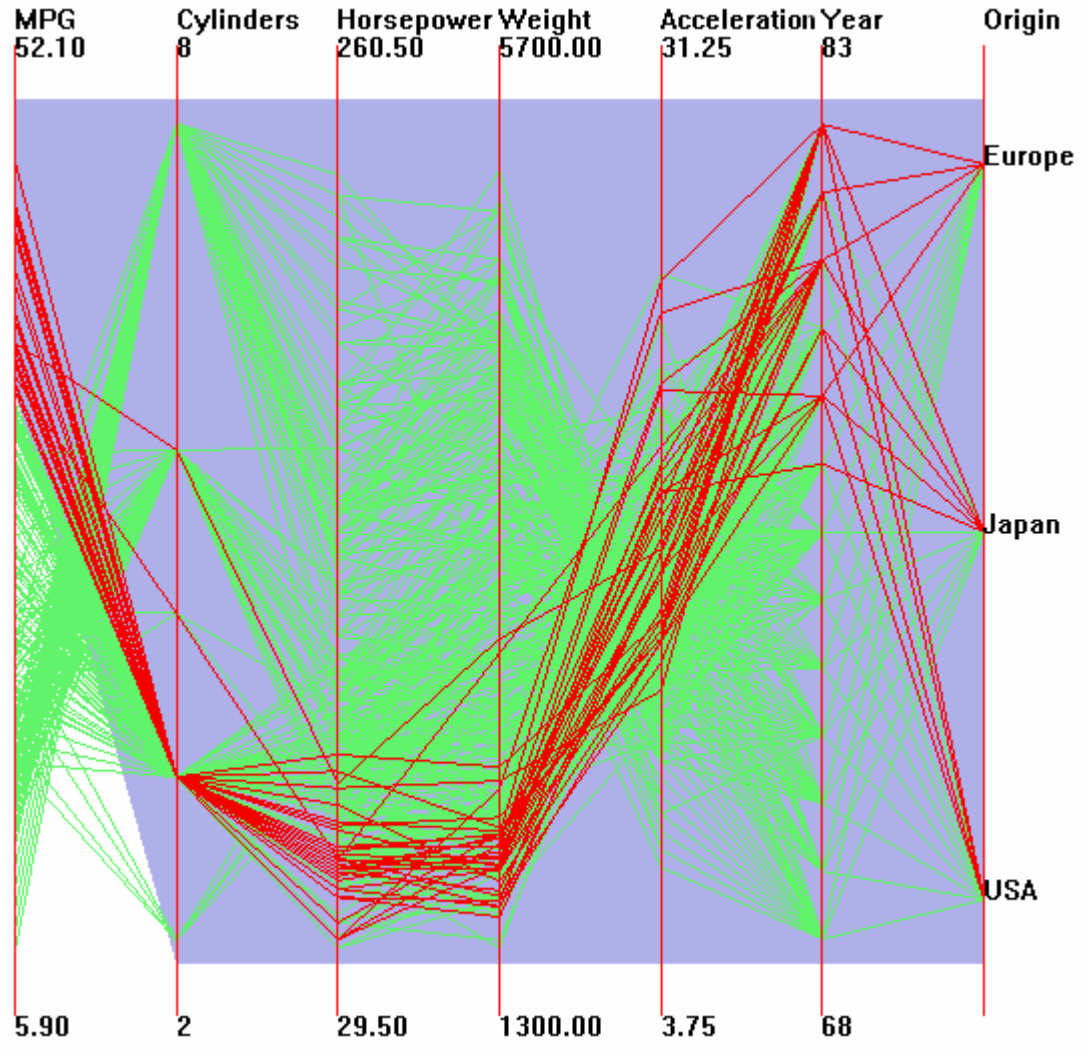

Figura 5.1: Aplicação da técnica Coordenadas Paralelas implementada pela ferramenta Xmdv, sobre a base de dados de automóveis

Na Figura 5.2, pode-se observar que a área de seleção não encobre mais os registros com 4 cilindros e, assim, os registros com 5 e 6 cilindros podem ser vistos separadamente.

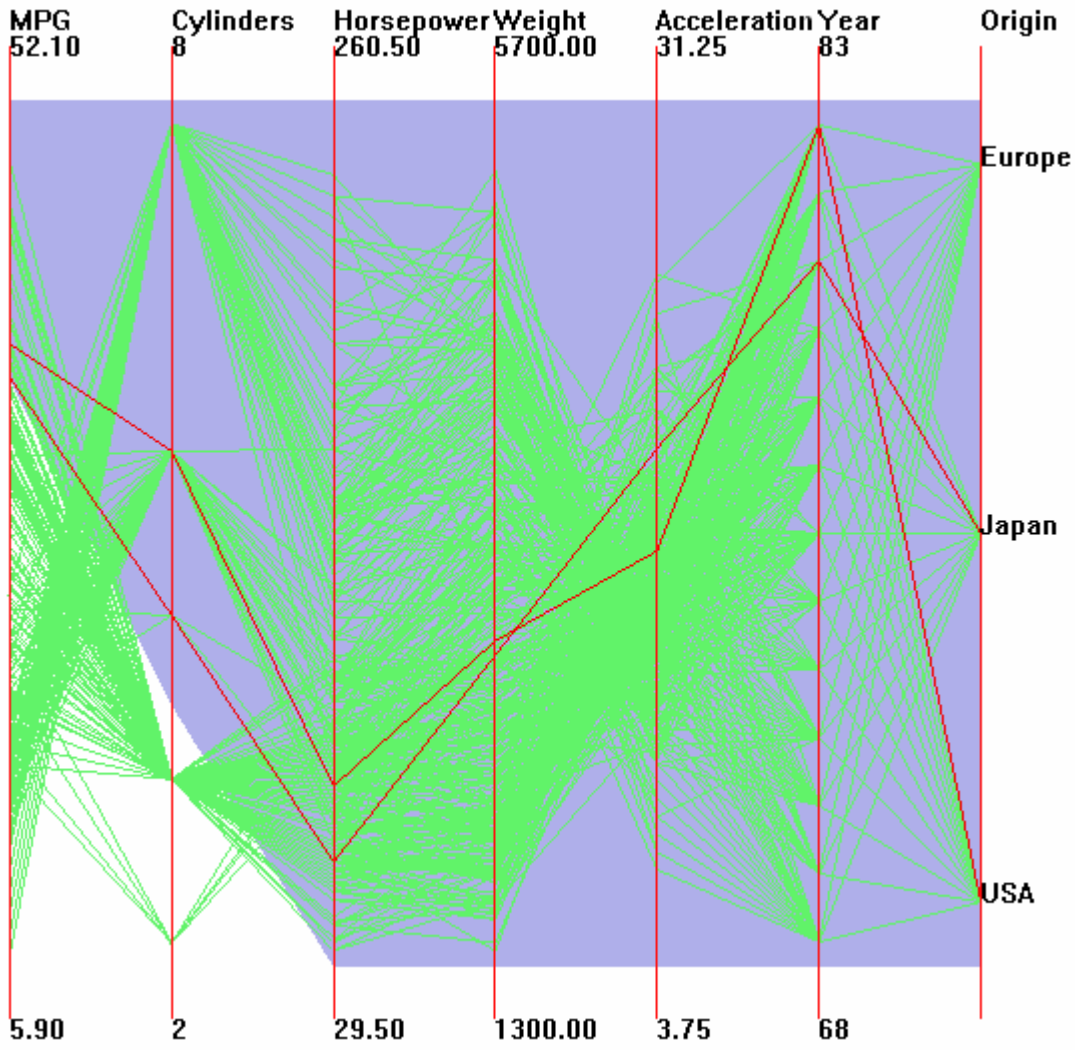

Figura 5.2: Aplicação da técnica Coordenadas Paralelas, implementada pela ferramenta Xmdv, sobre a base de dados de automóveis, destacando dois registros

A Figura 5.3 mostra a utilização da Matriz de Dispersão, implementada pela ferramenta Xmdv, sobre a mesma base de dados, na qual é possível observar a correlação entre cada par de atributos, gerando uma matriz 7x7.

Novamente, a área sombreada atrás dos registros representados equivale à mesma área de seleção definida quando foi utilizada a técnica coordenadas paralelas, mostrada na Figura 5.1. O mesmo ocorre para os registros em destaque que, no caso da matriz de dispersão, correspondem aos pontos com cor mais forte.

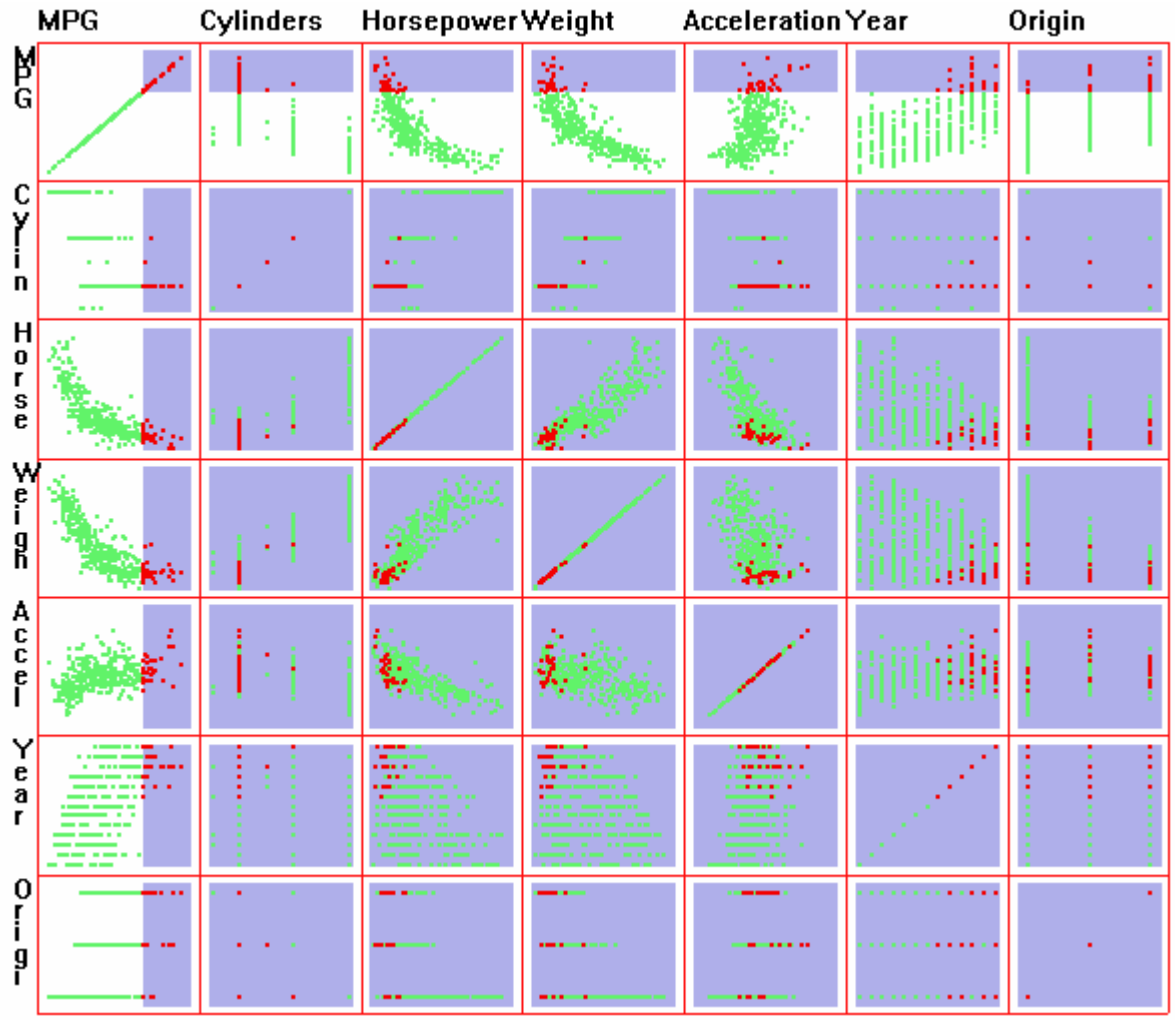

Figura 5.3: Aplicação da técnica Matriz de Dispersão, implementada pela ferramenta Xmdv, sobre a base de dados de automóveis

As características apresentadas pela base de dados de automóveis (dimensionalidade média, volume pequeno e atributos com valores quantitativos) permite a aplicação de técnicas iconográficas. A Figura 5.5 ilustra um exemplo de representação do conjunto de dados selecionados por meio da técnica *Star Glyphs*, implementada pela ferramenta Xmdv. Os ícones marcados por um círculo (feito manualmente) representam os mesmos carros com 5 e 6 cilindros, visualizados anteriormente na Figura 5.2. A ordenação dos atributos dos dados para as propriedades da estrela da *Star Glyph* é mostrada na Figura 5.4, cujo formato apresentado é equivalente à área de seleção dos atributos como já ilustrado pelas figuras anteriores.

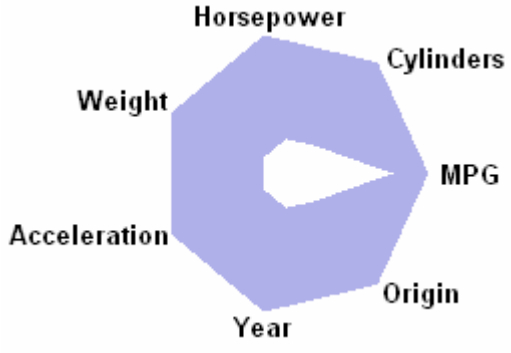

Figura 5.4: Ordenação das dimensões no ícone da Star Glyph da base de dados de automóveis

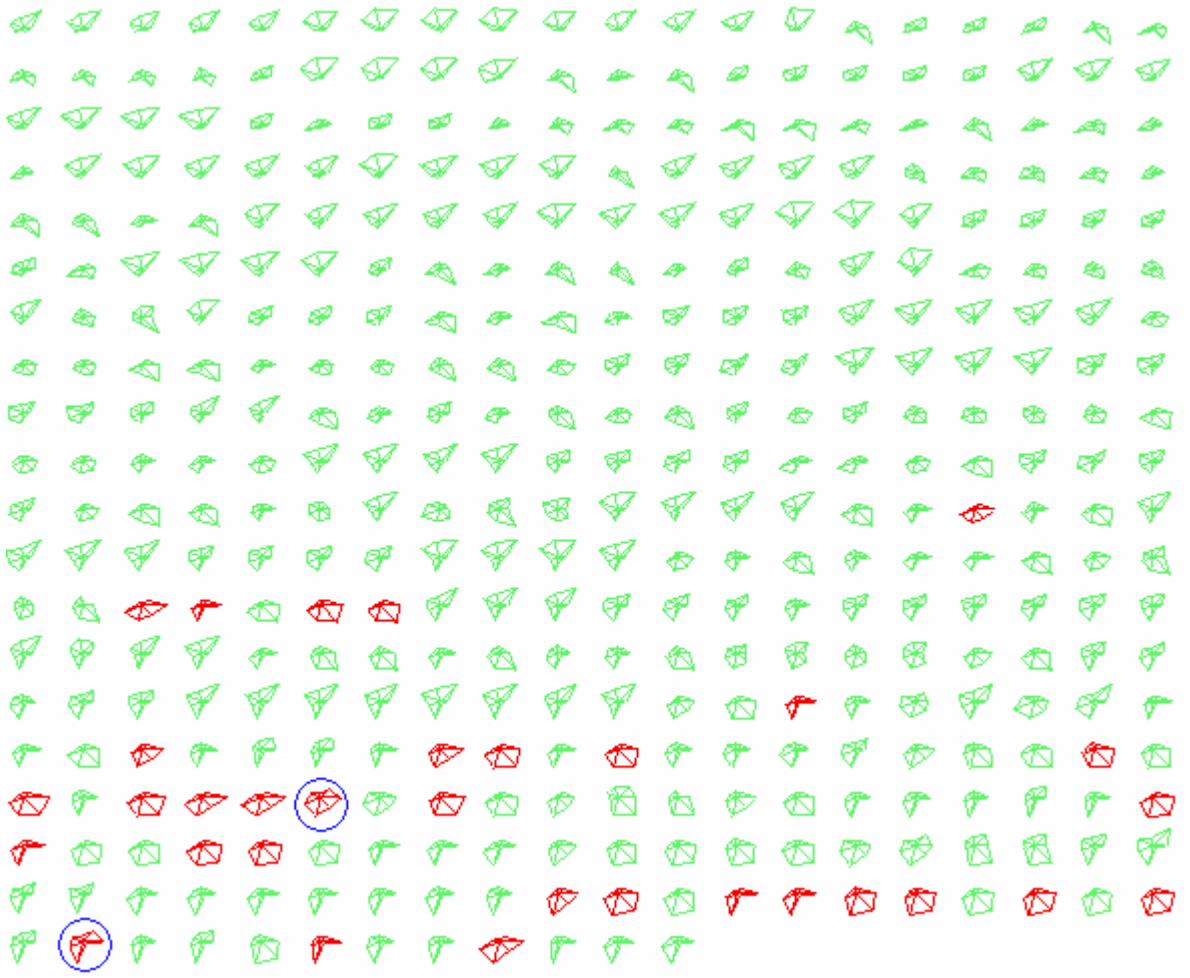

Figura 5.5: Aplicação da técnica Star Glyphs*, implementada pela ferramenta Xmdv, sobre a base de dados de automóveis*

c) Análise dos resultados

Os dados analisados possuem dimensionalidade média (7 atributos) e volume pequeno (392 registros), sendo os valores dos atributos em maioria do tipo quantitativo. Por isso, foram utilizadas as técnicas Coordenadas Paralelas, Matriz de Dispersão e *Star Glyphs*, fornecidas pela

ferramenta Xmdv, que permite a manipulação do posicionamento dos atributos manualmente ou por meio de algoritmos heurísticos.

Quanto ao parâmetro tipo de tarefa, a Tabela 5.3 sumariza as tarefas realizadas para analisar a base de dados de automóveis com as técnicas mencionadas. Na primeira coluna estão descritas as tarefas, de acordo com o parâmetro tipo de tarefa, enumeradas na Seção 4.2.2. Na segunda coluna estão as técnicas utilizadas para a respectiva tarefa e, na terceira coluna estão descritos exemplos de informações que podem ser obtidas pelas visualizações.

Tabela 5.3: Sumário das tarefas realizadas por meio das visualizações sobre a base de dados de automóveis

Tarefa	**Técnica**	**Exemplo de análise**
Visão geral	Coordenadas Paralelas Matriz de Dispersão	Visualização do relacionamento entre todas os atributos
Detecção de regras	Coordenadas Paralelas	Carros com alto valor de MPG possuem 4 cilindros e baixo peso
Outlier	Coordenadas Paralelas *Star Glyphs*	2 carros com 5 e 6 cilindros com alto valor de MPG e peso maior que o peso médio do grupo selecionado
Correlação entre os atributos	Coordenadas Paralelas Matriz de Dispersão	Relação entre MPG, cilindros e peso
Agrupamentos	Coordenadas Paralelas *Star Glyphs*	Seleção do grupo de carros com maior valor de MPG

5.3.2. Base de dados do Titanic

Esta base de dados possui informações sobre os passageiros do navio Titanic, disponível em <http://stats.math.uni-augsburg.de/Mondrian/>.

a) Características dos dados

A base descrita na Tabela 5.4 foi utilizada por Hofmann (2008) para apresentar a aplicação do gráfico de mosaicos, visto no Capítulo 4, na Seção 4.3.5. É composta por quatro atributos qualitativos ou categóricos. O atributo *Class* representa a classe social do passageiro, distinto entre os valores *fisrt*, *second*, *third* e *crew* (primeira, segunda, terceira e tripulação); o atributo *Age* indica a faixa etária dos indivíduos, dividida entre criança (*child*) e adulto (*adult*); o atributo *Sex* informa o gênero da pessoa; e o atributo *Survived* classifica os passageiros como sobreviventes e não sobreviventes. Por apresentar dados categóricos, técnicas da classe hierárquica e baseada em grafos podem ser utilizadas para representar esta base de dados.

Tabela 5.4: Base de dados dos passageiros do navio Titanic

Parâmetros	**Valor**	**Características**
Número de dimensões	4	dimensionalidade baixa
Quantidade de registros	2201	volume médio
Descrição dos atributos	Classe (*Class*)	qualitativo ordinal
	Idade (*Age*)	qualitativo ordinal
	Sexo (*Sex*)	qualitativo nominal
	Sobrevivente (*Survived*)	qualitativo nominal

b) Aplicação de técnicas de visualização

Para visualizar esta base de dados, foi utilizada a ferramenta Mondrian que, assim como a Xmdv, possibilita a troca de informações entre as visualizações (*linked view*). Esta ferramenta foi utilizada por implementar técnicas que permitem a visualização de dados qualitativos, como o gráfico de mosaicos.

Outra ferramenta utilizada para visualizar esta base de dados foi a Treemap, cujo nome já indica que implementa a técnica *Treemap*. Por meio desta ferramenta foi possível explorar os dados sob a perspectiva de sua estrutura hierárquica.

A Figura 5.6 ilustra a aplicação da ferramenta Mondrian. Nesta figura nota-se o destaque para um subconjunto de dados que reflete-se entre os gráficos de barra e o gráfico de mosaico, em que é analisada a classe tripulação (*crew*), composta somente por adultos, sendo grande maioria do sexo masculino e que também corresponde a maior taxa de não-sobreviventes dessa classe. No canto inferior direito dessa ilustração está a orientação para a leitura da alocação dos atributos no gráfico de mosaico.

Figura 5.6: Aplicação do Gráfico de Mosaicos, implementada pela ferramenta Mondrian, sobre a base de dados dos passageiros do Titanic

As Figuras 5.7 e 5.8 mostram o emprego da técnica *Treemap* implementada pelo software Treemap desenvolvido pelo grupo HCIL. Uma das informações que pode-se observar é o fato de todas as crianças da primeira e segunda classes terem sobrevivido. Outro fato pode ser notado confrontando a Figura 5.7, que mostra a relação de sobreviventes, com a Figura 5.8, que mostra a proporção de homens e mulheres: a taxa de sobreviventes é maior entre as mulheres do que entre os homens, e essa taxa aumenta conforme a classe social. Isto também foi verificado por Hofmann (2008), conforme discutido na Seção 4.3.5.

Figura 5.7: Aplicação de Treemap*, implementada pela ferramenta Treemap, sobre a base de dados do Titanic, destacando sobreviventes e não sobreviventes*

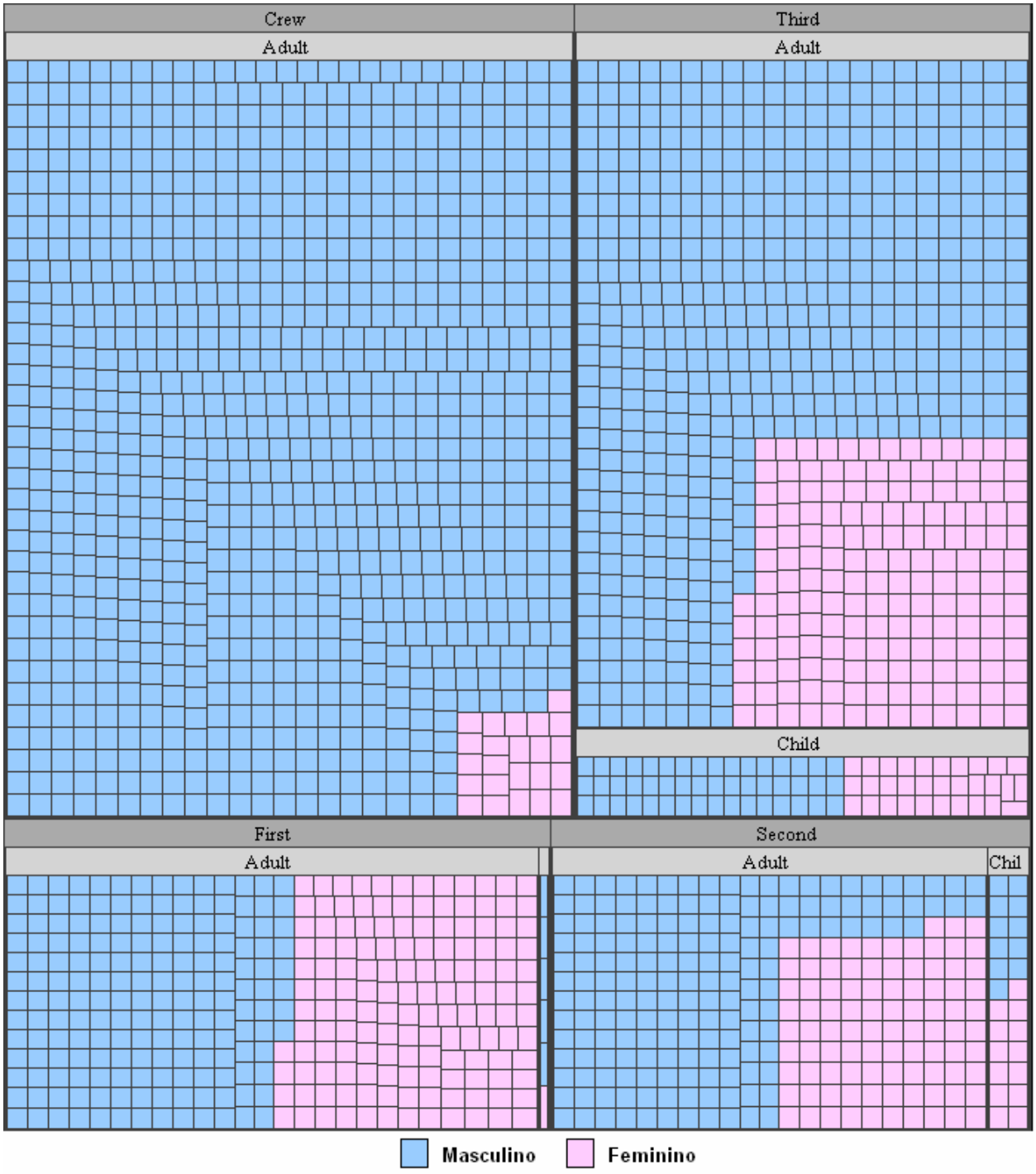

Figura 5.8: Aplicação de Treemap*, implementada pela ferramenta Treemap, sobre a base de dados do Titanic, destacando homens e mulheres*

Hofmann (2008) apresenta a visualização dessa base de dados pela estrutura de árvore. O autor afirma que quando a árvore representa uma tabela de eventos, como é o caso da base de dados do Titanic, todas as ramificações possuem a mesma profundidade e todos os nós que estão em um mesmo nível representam o mesmo atributo, como pode ser visto na Figura 5.9.

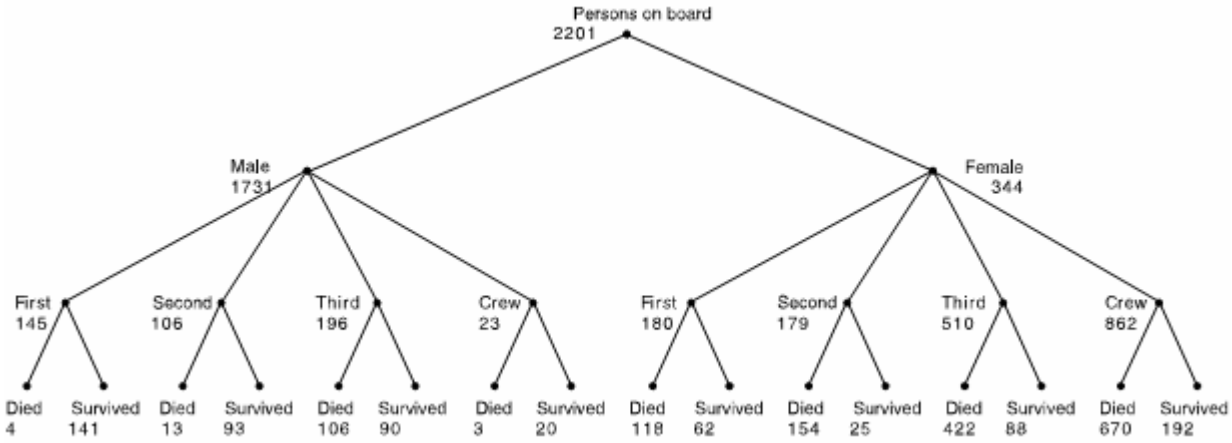

Figura 5.9: Aplicação de grafos sobre a base de dados do Titanic (Hofmann, 2008; p.636)

c) Análise dos resultados

A base de dados analisada é formada por valores qualitativos, que correspondem a determinadas categorias de passageiros do Titanic. Outras características conforme os parâmetros definidos são a dimensionalidade baixa (4 atributos) e volume médio (2201 registros). Dessa forma, foram aplicadas técnicas da classe hierárquica ou baseada em grafos, que são: Gráfico de Mosaicos, da ferramenta Mondrian e *Treemaps*, utilizando a ferramenta Treemap e um exemplo de Grafos dado por Hofmann (2008).

Mondrian obedece a ordem dos atributos em que está organizada na base de dados, não permitindo a manipulação do posicionamento dos atributos diretamente pelos gráficos. Já pela ferramenta Treemap é possível alterar a ordem de hierarquia dos atributos, que no caso das Figura 5.7 e Figura 5.8 estão organizadas pela classe social dos passageiros distintos por sua vez entre adultos e crianças.

A Tabela 5.5 sumariza as tarefas realizadas para analisar a base de dados do Titanic utilizando essas técnicas. Na primeira coluna estão descritas as tarefas, de acordo com o parâmetro tipo de tarefa, enumeradas na Seção 4.2.2. Na segunda coluna estão as técnicas utilizadas para a respectiva tarefa e, na terceira coluna estão descritos exemplos de informações que podem ser obtidas pelas visualizações.

Tabela 5.5: Sumário das tarefas realizadas por meio das visualizações sobre a base de dados do Titanic

Tarefa	**Técnica**	**Exemplo de análise**
Visão geral	Grafos	Todas as categorias são visíveis pelos nós da árvore
Detecção de padrões	Gráfico de Mosaicos *Treemaps*	Taxa de mulheres sobreviventes é maior que a de homens, e esse valor aumenta conforme a classe social
Outlier	Sem exemplos	Sem exemplos
Correlação entre os atributos	Gráfico de Mosaicos *Treemaps*	Relação entre classe social e taxa de sobreviventes
Agrupamentos	*Treemaps*	Identificação de grupos baseado em cores

5.3.3. Base de dados do Vestibular

Refere-se aos dados da pontuação obtida pelos candidatos ao curso de Medicina pelo vestibular da Universidade Estadual de Maringá (UEM), ocorrido em dezembro de 2009, disponível em <http://www.cvu.uem.br/>.

a) Características dos dados

A Tabela 5.6 descreve as principais características notadas sobre esse conjunto de dados, com dimensionalidade média (8 atributos) e volume médio, com 1572 registros. Os atributos possuem valores quantitativos contínuos, representando as notas obtidas pelos candidatos para cada tipo de prova, e quantitativo discreto para a classificação, com exceção do atributo situação, qualitativa nominal, que denomina candidatos aprovados, reprovados e em lista de espera.

Tabela 5.6: Base de dados dos candidatos ao curso de Medicina do vestibular da UEM de 2009

Parâmetros	**Valor**	**Características**
Número de dimensões	8	dimensionalidade média
Quantidade de registros	1572	volume médio
Descrição dos atributos	nota prova redação (RED)	quantitativo contínuo
	nota língua portuguesa (PORT)	quantitativo contínuo
	nota língua estrangeira (LE)	quantitativo contínuo
	nota biologia (BIO)	quantitativo contínuo
	nota química (QUI)	quantitativo contínuo
	nota final (TOTAL)	quantitativo contínuo
	classificação (CLAS)	quantitativo discreto
	situação (SIT)	qualitativo nominal

b) Aplicação de técnicas de visualização

Pelas características apresentadas por essa base de dados, pode-se optar, a princípio, por técnicas geométricas. Se houver um processo de redução de dados, por exemplo, filtrando os candidatos aprovados, pode-se aplicar também técnicas iconográficas.

Diante do formato dos dados, foi utilizada a ferramenta CASSATT para explorar toda a base de dados, correspondente a todos os candidatos inscritos para o vestibular de medicina, e a ferramenta GAUGUIN para visualizar o subconjunto de registros daqueles que foram aprovados.

A Figura 5.10 ilustra a representação dessa base de dados pela técnica Coordenadas Paralelas implementada pela ferramenta CASSATT. Nessa figura estão em destaque os candidatos com situação de aprovado (indicado pelo valor AP do atributo situação – SIT).

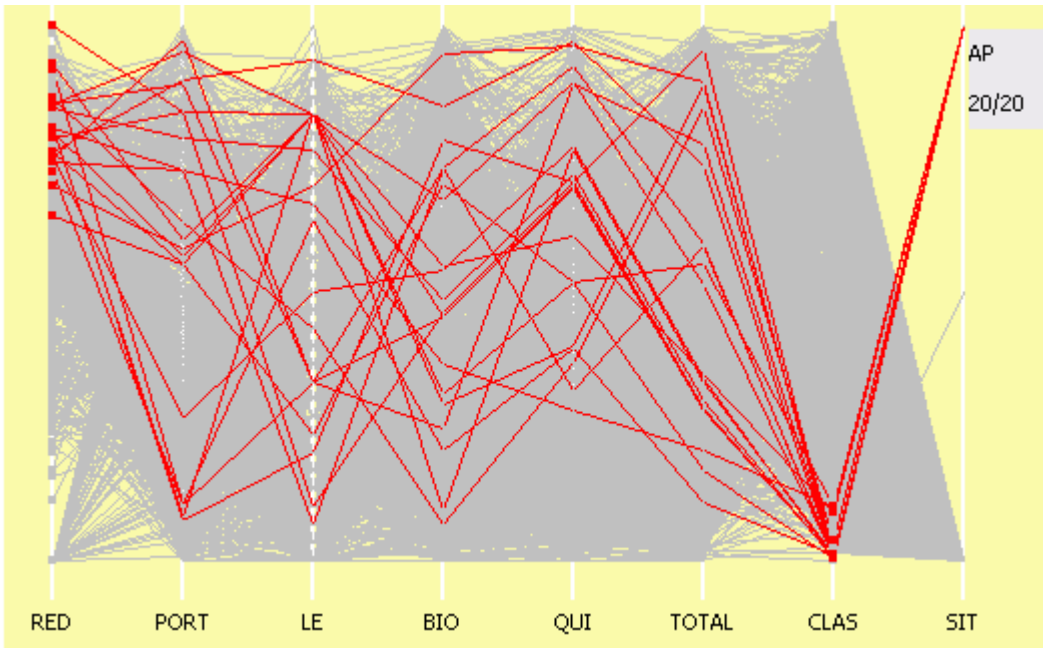

Figura 5.10: Aplicação de Coordenadas Paralelas, implementada pela ferramenta CASSATT, destacando os candidatos aprovados do vestibular de verão da UEM de 2009

A Figura 5.11 mostra o processo de ligação (*linked view*) entre diferentes tipos de visualização, implementada pela ferramenta CASSATT, em que os registros selecionados por meio das coordenadas paralelas (Figura 5.10) podem ser visualizados pelo diagrama de caixas construído para cada atributo.

É possível notar em observação conjunta das Figuras 5.10, 5.11 e Tabela 5.7 que os candidatos aprovados destacaram-se na prova de redação e a maioria obteve uma boa pontuação também na prova de química. Este fato pode ser observado na Figura 5.10, pela região mais alta

dos eixos que são tocados pelas linhas poligonais da coordenada paralela; na Figura 5.11, pela observação dos diagramas de caixa para cada atributo na qual estão em destaque os registros dos candidatos aprovados; e na Tabela 5.7, são informados os valores da média e mediana fornecidos pela ferramenta CASSATT.

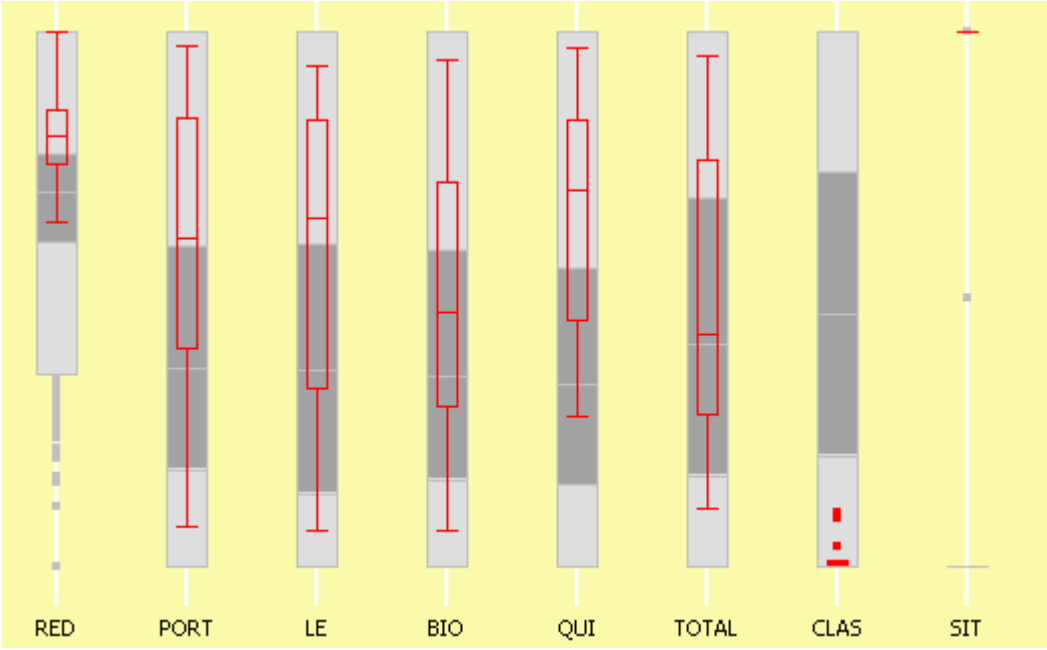

Figura 5.11: Aplicação do Diagrama de Caixas, implementada pela ferramenta CASSATT, destacando os candidatos aprovados do vestibular de verão da UEM de 2009

Tabela 5.7: Valores da média e mediana das notas dos candidatos aprovados no vestibular para medicina informados pela ferramenta CASSATT

	RED	PORT	LE	BIO	QUI	TOTAL	CLAS
Média	70,8	54,42	12,33	208,48	75,52	453,52	700,93
Mediana	76,0	51,0	12,0	186,0	69,0	433,5	698,5

A Figura 5.12 mostra as Coordenadas Paralelas com destaque aos candidatos reprovados (indicado pelo valor RP do atributo situação – SIT). Os registros referentes aos candidatos reprovados tiveram o valor do atributo situação colocados em 0.0, como indicado na figura, pela qual também pode-se observar que o principal motivo da reprovação é a pontuação baixa na prova de redação.

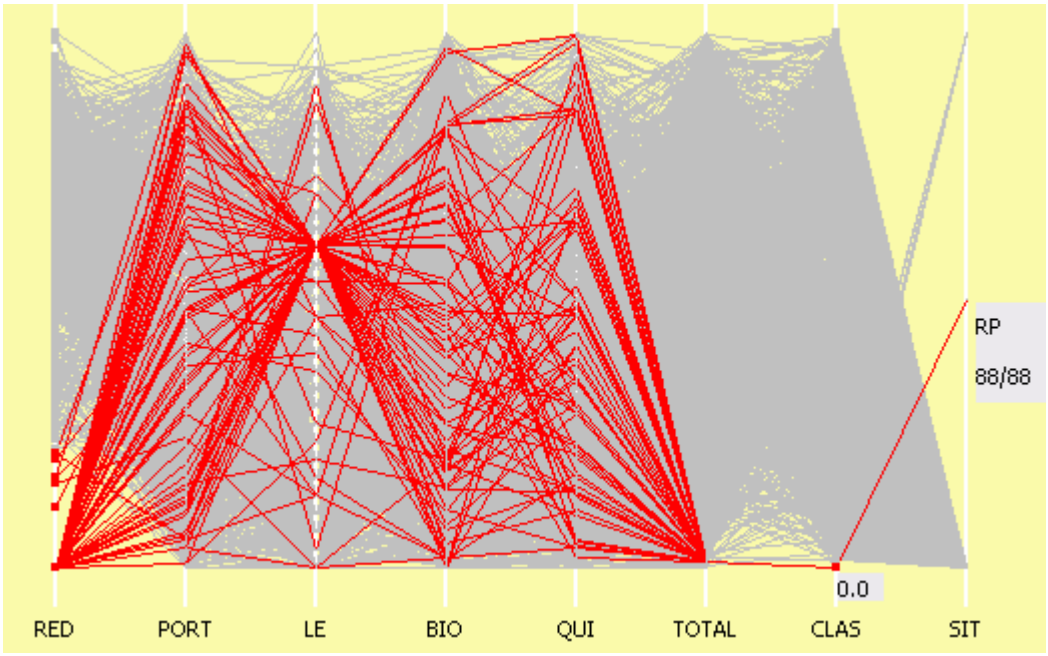

Figura 5.12: Aplicação das Coordenadas Paralelas, implementada pela ferramenta CASSATT, destacando os candidatos reprovados do vestibular de verão da UEM de 2009

Os registros correspondentes aos candidatos aprovados também podem ser visualizados por uma técnica de visualização da classe iconográfica, pois o volume de dados reduz-se a um total de 20 elementos. A ferramenta GAUGUIN provê, dentre outras técnicas, a *Star Glyph* que foi utilizada para visualizar de outra forma esse conjunto de dados.

A Figura 5.13 mostra o mapeamento dos atributos para o ícone da *Star Glyph*, e a Figura 5.14 ilustra os 20 registros dos candidatos aprovados.

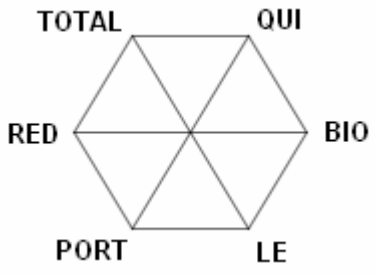

Figura 5.13: Ordem dos atributos da base de dados do vestibular 2009 para o ícone da técnica Star Glyphs da ferramenta GAUGUIN

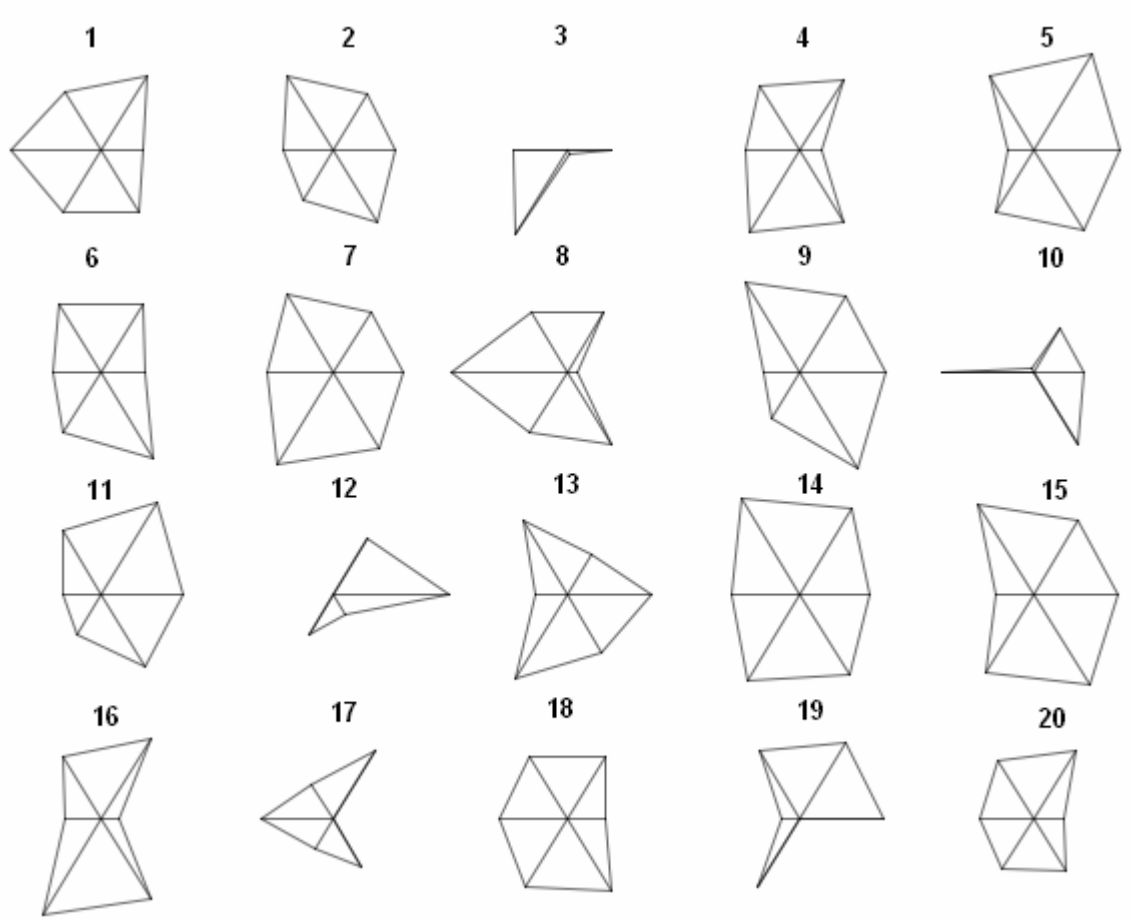

Figura 5.14: Aplicação da técnica Star Glyph, implementada pela ferramenta GAUGUIN, sobre a base de dados do vestibular de verão da UEM de 2009, destacando os candidatos aprovados

Uma curiosidade é o formato discrepante apresentado pelos ícones 3, 10, 12 e 17, em relação aos demais registros. Isto deve-se ao fato de que esses candidatos foram aprovados por meio da política de cotas adotada pela universidade para a realização do vestibular.

c) Análise dos resultados

A base de dados analisada possui dimensionalidade média (8 atributos) e volume médio (1572 registros). Exceto pelo atributo situação, do tipo qualitativo nominal, o restante dos atributos são do tipo quantitativo. Todo o volume de dados foi representado pelas Coordenadas Paralelas, dessa vez utilizando a implementação da ferramenta CASSATT, pois esta técnica pode representar ao mesmo tempo valores quantitativos e qualitativos, como foi demonstrado anteriormente, no exemplo da base de dados dos automóveis. Para um grupo menor de dados, formado por registros dos candidatos aprovados, foi aplicada a técnica iconográfica *Star Glyph*, implementada na ferramenta GAUGUIN.

CASSAT permite a manipulação manual dos eixos das Coordenadas Paralelas, mas o posicionamento dos atributos, no caso dessa base de dados, não interfere na compreensão dos resultados, pois são quase todas variáveis independentes, que representam as notas individuais obtidas pelos candidatos. GAUGUIN permite que os ícones da *Star Glyphs* possam ser ordenados em relação a um dos atributos da base de dados.

A Tabela 5.8 resume as tarefas que puderam ser realizadas utilizando as técnicas de visualização sobre os dados referentes aos candidatos ao curso de medicina do vestibular de verão da Universidade Estadual de Maringá, ocorrido no mês de dezembro de 2009.

Tabela 5.8: Sumário das tarefas realizadas por meio das visualizações sobre a base de dados do vestibular de verão da UEM do ano de 2009

Tarefa	**Técnica**	**Exemplo de análise**
Visão geral	Coordenadas Paralelas	Visualização de todos os registros da base de dados, em que é possível selecionar os que são de interesse
Detecção de padrões	Coordenadas Paralelas Diagrama de caixas	A prova de redação é uma das avaliações decisivas para a classificação do candidato
Outlier	*Star Glyph*	Quatro candidatos aprovados que não obtiveram média semelhante aos demais
Correlação entre os atributos	Sem exemplo	Sem exemplo
Agrupamentos	Coordenadas Paralelas	Identificação do grupo de candidatos aprovados e reprovados

5.3.4. Base de dados Depressão Pós-Parto

Os dados referem-se a um estudo de depressão pós-parto, feito por uma residente em Psiquiatria da UEM em 2006.

a) Características dos dados

Os dados descritos na Tabela 5.9 possuem valores do tipo quantitativo e qualitativo, típicas de questionários com alternativas fechadas, em que as entrevistadas responderam questões relacionadas a: Idade; Escolaridade: primeiro grau incompleto, primeiro grau completo, segundo grau incompleto, segundo grau completo, superior incompleto, APAE; Estado civil: solteira, casada, relação estável; Depressão familiar: sim, não; Religião: Católica, Evangélica, nenhuma; Atividade: estudante, do lar, emprego registrado (CLT), desempregada; Renda familiar.

Tabela 5.9: Base de dados de estudo sobre depressão pós-parto

Parâmetros	**Valor**	**Características**
Número de dimensões	8	dimensionalidade média
Quantidade de registros	68	volume pequeno
Descrição dos atributos	Idade	quantitativo discreto
	Escolaridade	qualitativo ordinal
	Estado_Civil	qualitativo nominal
	Depre_Familiar	qualitativo nominal
	Religião	qualitativo nominal
	Atividade	qualitativo nominal
	Renda	quantitativo contínuo
	Escore	quantitativo discreto

b) Aplicação de técnicas de visualização

O atributo Escore refere-se à pontuação alcançada pelas pacientes que preencheram o questionário. Quanto maior a pontuação, maior a probabilidade de tratar-se de um caso de depressão pós-parto. A pontuação máxima que pode-se alcançar é 30 e a mínima é 0. Pela literatura de base da área, foi considerado casos sugestivos de depressão pós-parto aqueles com pontuação igual ou maior a 12. Assim, foram destacados os registros cujo valor do atributo Escore estava nessa faixa.

A aplicação da visualização, neste exemplo, está voltada para a análise da correlação entre os atributos, que foi possível com as Coordenadas Paralelas, pois esta técnica permite a manipulação de tipos híbridos de dados (que possui tipos quantitativos e qualitativos). A Matriz de Dispersão não é aplicável a dados categóricos (qualitativos), logo, sua utilização foi descartada para esse conjunto de dados para a verificação da correlação. Técnicas iconográficas podem ser utilizadas para visualizar a relação entre os atributos, mas é uma classe de técnicas em que a correlação entre os dados é mais difícil de ser detectada (Rabelo, 2007).

A Figura 5.15 mostra a aplicação das Coordenadas Paralelas para todas os atributos, utilizando a ferramenta CASSATT, que exibe os registros com valor qualitativo por meio de pontos cujos tamanhos variam conforme a quantidade de registros correspondentes.

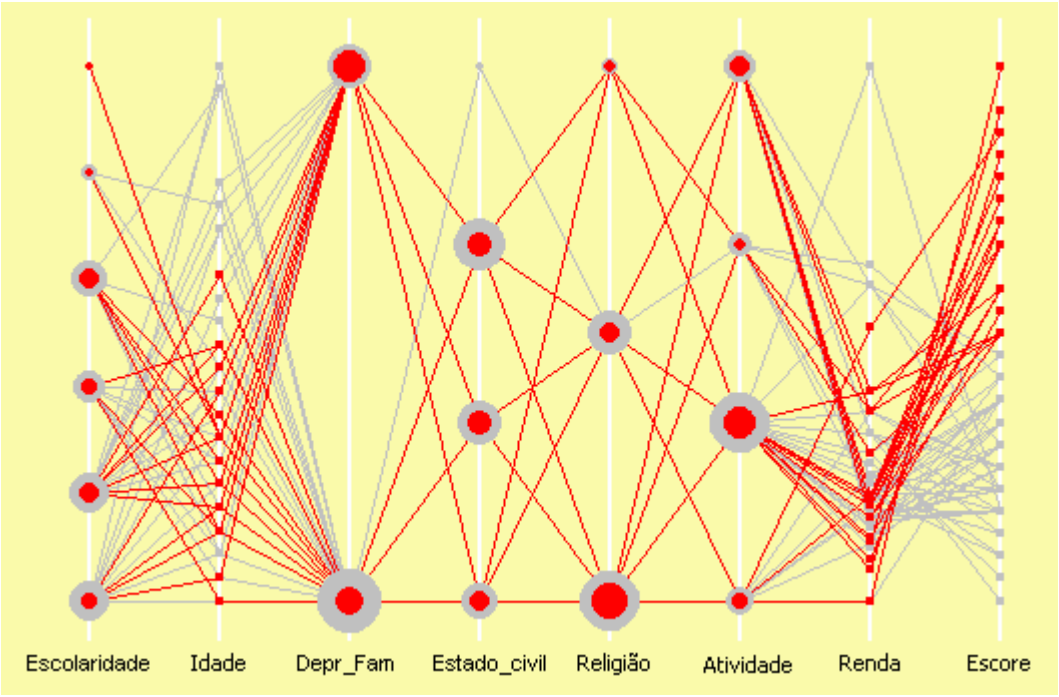

Figura 5.15: Aplicação das Coordenadas Paralelas, implementada pela ferramenta CASSATT, para a análise dos atributos da base de dados sobre depressão pós-parto

Devido a presença de vários atributos qualitativos (com valores variando entre duas a quatro alternativas), as linhas poligonais das coordenadas paralelas acabam sobrepondo-se, por isso, os atributos foram divididos em duas diferentes visualizações, ilustradas pela Figura 5.16 e Figura 5.17

A Figura 5.16 exibe os atributos Religião, Idade, Escolaridade, Depre_Familiar e Escore. Pode-se perceber que, para essa amostra de dados, trata-se de mulheres jovens e a maioria declara-se católica.

A Figura 5.17 exibe os atributos Idade, Atividade, Renda, Estado_civil e Escore. Nessa figura percebe-se, novamente para essa amostra de dados, que a maioria das entrevistadas é dona de casa.

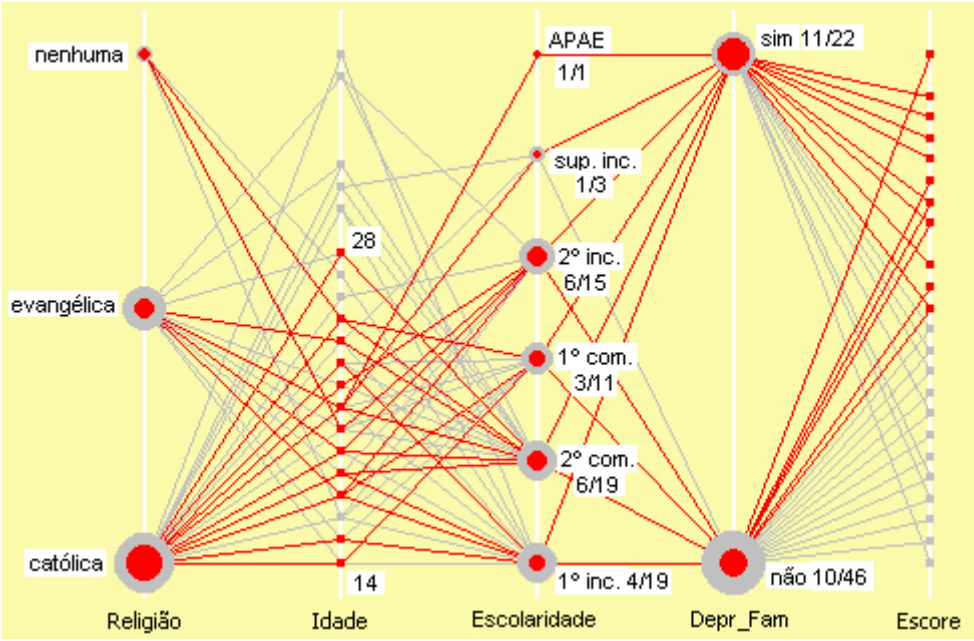

Figura 5.16: Aplicação das Coordenadas Paralelas, para a análise do primeiro conjunto de atributos da base de dados sobre depressão pós-parto

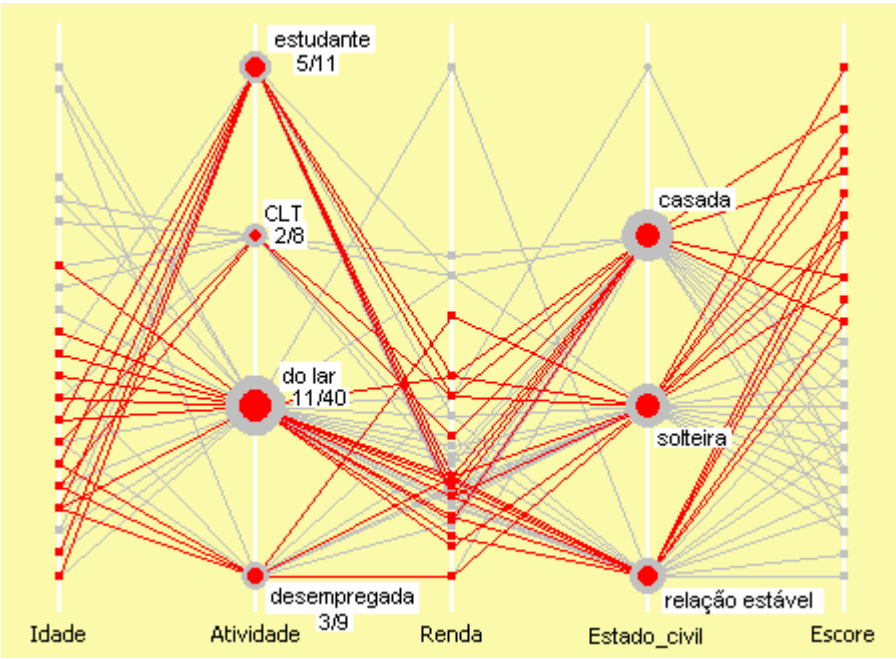

Figura 5.17: Aplicação das Coordenadas Paralelas, para a análise do segundo conjunto de atributos da base de dados sobre depressão pós-parto

c) Análise dos resultados

A base de dados possui dimensionalidade média (8 atributos: 5 qualitativos e 3 quantitativos) e volume pequeno (68 registros). Pela característica híbrida dos dados, a técnica eleita para construir esse exemplo foi a Coordenada Paralela, implementada pela ferramenta CASSATT. Houve o problema da sobreposição dos registros na visualização de todos os atributos, por isso, foram geradas visualizações para dois grupos de atributos.

Uma pessoa leiga pode identificar por meio das visualizações as informações mais evidentes, como a faixa de idade ou a média de renda familiar das entrevistadas, calculada pela ferramenta. Porém, só um especialista do domínio é capaz de ponderar a relevância de um atributo para o desenvolvimento de uma depressão pós-parto, sendo as visualizações um meio pelo qual o analista pode apoiar-se para explorar os dados, selecionando atributos de interesse. A Tabela 5.10 sumariza as tarefas realizadas por meio da visualização dessa base de dados.

Tabela 5.10: Sumário das tarefas realizadas por meio das visualizações sobre a base de dados sobre depressão pós-parto

Tarefa	**Técnica**	**Exemplo de análise**
Visão geral	Coordenadas Paralelas	Visualização de todos os registros da base de dados, em que é possível selecionar os casos possíveis de depressão pós-parto
Detecção de padrões	Coordenadas Paralelas	Para a amostra analisada, mulheres com propensão a desenvolver depressão pós-parto são jovens, na faixa etária de 14 a 28 anos
Outlier	Sem exemplo	Sem exemplo
Correlação entre os atributos	Coordenadas Paralelas	As mulheres entrevistadas possuem renda familiar média de R$ 607,27 e são em sua maioria jovens e donas de casa
Agrupamentos	Sem exemplo	Sem exemplo

5.4. Considerações Finais

Uma dificuldade que pode ser encontrada durante o processo de extração de conhecimento é a preparação dos dados que alimentarão a ferramenta para gerar a visualização, pois cada software possui uma implementação particular e, consequentemente, aceita um padrão de entrada específico.

Vale ressaltar que o sucesso da exploração visual depende do quão significativo são os dados para quem está analisando-os. A familiaridade do analista com os dados eleva os benefícios da visualização para a aquisição do conhecimento, permitindo ao analista elaborar e experimentar novas hipóteses ou detectar novos padrões e características importantes. Um exemplo é a análise feita sobre a base de dados de depressão pós-parto, descrita na Seção 5.3.4, em que o conhecimento do especialista é fundamental para direcionar a exploração visual dos dados.

6. Diretrizes para escolher técnicas de visualização

6.1. Considerações iniciais

Neste capítulo é apresentada a fase final da Teoria Fundamentada em Dados correspondente ao procedimento de formulação da teoria que emerge do resultado das etapas de análise e codificação dos dados, sendo acompanhada de um diagrama desenvolvido para ajudar no entendimento da mesma.

Assim, o processo de análise e codificação apresentado nos Capítulos 4 e 5 teve como resultado a formulação das diretrizes para a escolha de técnicas de visualização. Essas diretrizes são baseadas em uma análise final sobre a associação de cada parâmetro identificado com a classe de técnica de visualização.

De modo geral, quando pretende-se utilizar técnicas de visualização em um processo de descoberta de conhecimento, o procedimento a ser seguido na decisão de quais técnicas utilizar segue o cenário ilustrado pela Figura 6.1.

Este cenário descreve que, a princípio, há uma base de dados da qual fica a critério do analista definir se haverá um subconjunto de dados destinados para a análise. Definido este conjunto de dados, o próximo passo é observar as características pertinentes aos dados, tais como tipo, dimensionalidade e volume. O parâmetro posição dos atributos é levado em consideração quando a técnica de visualização utilizada requer que a exploração dos dados seja feita por meio da manipulação dos elementos do gráfico que representam os atributos.

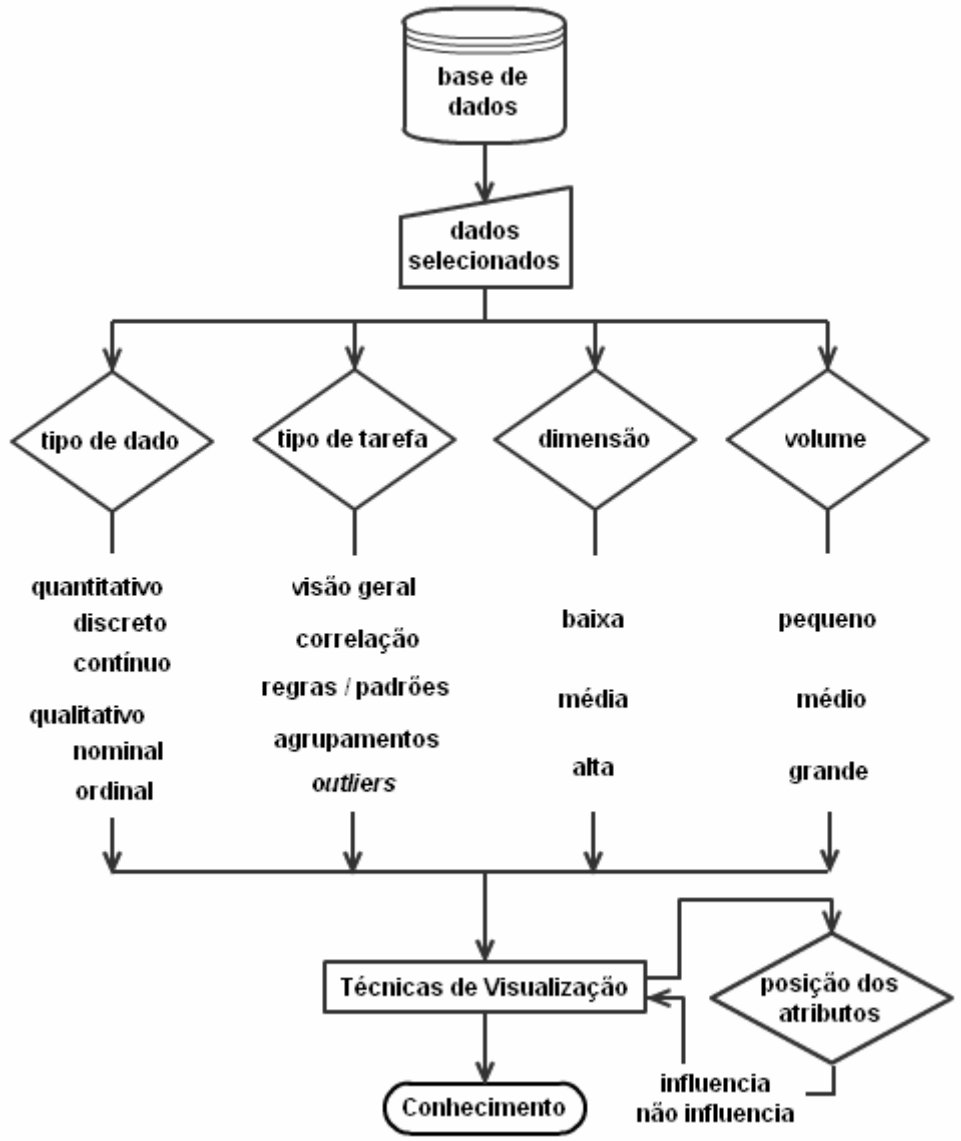

Figura 6.1: Cenário do procedimento para o emprego de técnicas de visualização no processo de extração de conhecimento

O tipo de dado é um fator determinante para a escolha de técnicas de visualização, em razão de que existem técnicas de visualização projetadas para exibir somente dados qualitativos, quantitativos, hierárquicos e também para formatos específicos como mídia, códigos-fonte de programas, documentos, etc. No cenário descrito, é considerado somente o aspecto da natureza do dado, se o tipo é qualitativo ou quantitativo, por ser este o mais encontrado durante a pesquisa sobre as técnicas de visualização estudadas no Capítulo 4.

A dimensionalidade dos dados é avaliada de acordo com o número de atributos. Neste trabalho convencionou-se que: dados com até 4 atributos possuem dimensionalidade baixa; com 5 a 9 atributos, dimensionalidade média; acima de 10 atributos, dimensionalidade alta.

Da mesma forma, o volume dos dados também foi convencionado em categorias que

distingue o volume de dados em pequeno, médio e grande. Assim, um conjunto de dados com quantidade de registros variando na ordem de 10^1 a 10^2 itens, é considerado como volume pequeno; se estiver na ordem de 10^3 a 10^5, é considerado como volume médio; na ordem de 10^6 a 10^7, é considerado como volume grande.

As tarefas que o usuário realizará para a exploração dos dados também devem ser consideradas, as quais basicamente são: visão geral dos dados; verificação de correlação entre os atributos; identificação de regras ou padrões novos; análise de agrupamentos e detecção de *outliers.*

Além do tipo de dado, tipo de tarefa, dimensionalidade e volume, outro parâmetro a ser considerado é o posicionamento dos atributos no gráfico. Dependendo da técnica de visualização utilizada, a ordem na qual os atributos estão posicionados pode influenciar no resultado da exploração dos dados.

Descrito o cenário ilustrado na Figura 6.1, na sequência os parâmetros são analisados no contexto das categorias de técnicas de visualização: 1D a 3D, iconográficas, geométricas, orientadas a pixel e hierárquicas, como ilustra a Figura 6.2.

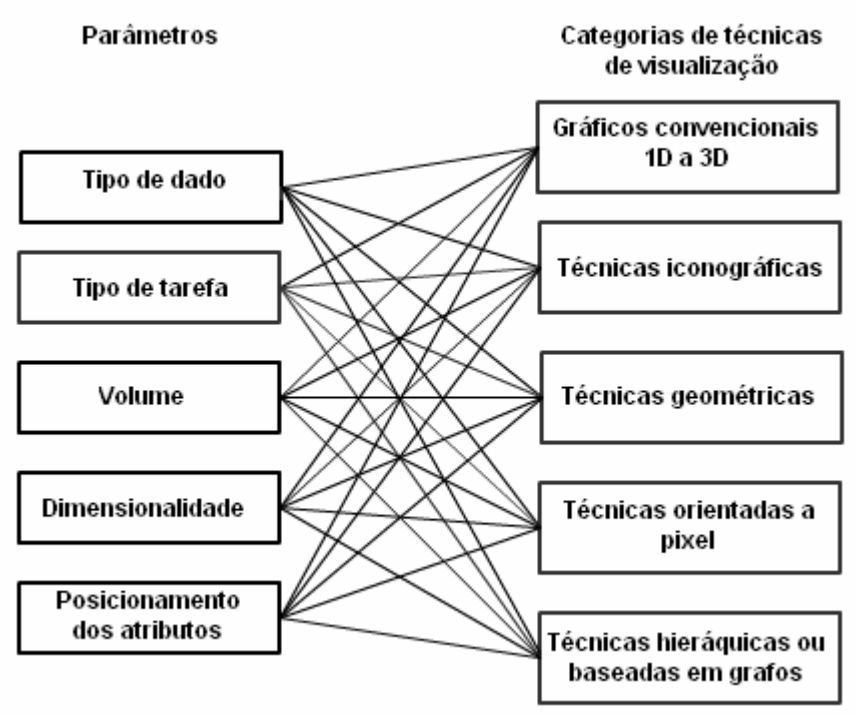

Figura 6.2: Associação entre os parâmetros identificados e as classes de técnicas de visualização

6.2. Tipo de dado

O tipo de dado é um fator determinante para a escolha de técnicas de visualização. Keim (2002) e Shneiderman (1996) classificam as técnicas de visualização pelo tipo de dado, que pode ser: de uma, duas, três ou *n* dimensões; quantitativo ou qualitativo; documentos ou hipertextos; algoritmos ou softwares; hierárquicos ou descritos por grafos. A Tabela 6.1 sumariza o tipo de dado mais indicado para ser visualizado por cada técnica considerada neste trabalho, levando em consideração a natureza de domínio do dado (como indicado pela Tabela 4.8, p.62).

Tabela 6.1: Técnicas de visualização e o tipo de dado que melhor podem representar

Categoria	**Técnicas**	**Dados quantitativos**		**Dados qualitativos**	
		Contínuos	Discretos	Nominais	Ordinais
1-D a 3-D	Histograma	X	X	X	X
	Diagrama de caixa	X			
	Gráfico de dispersão	X			
	Gráfico de contorno	X	X		
Iconográficas	Faces de Chernoff	X	X		
	Star Glyphs	X	X		
	Figuras de Aresta	X	X		
Geométricas	Matriz de dispersão	X			
	Coordenadas Paralelas	X	X	X	X
Orientadas a pixel	*Query-dependent techniques*	X	X		
	Query-independent techniques	X	X		
Hierárquicas*	Grafos	X	X	X	X
	Cone trees	X	X	X	X
	Treemap	X	X	X	X
	Gráfico de mosaico	X	X	X	X
	Empilhamento de dimensões	X	X	X	X

* Os dados devem possuir uma estrutura de relacionamento que possa ser representada por esta classe de visualização.

Técnicas da classe 1-D a 3-D, que representam de um a três atributos, podem ser utilizadas para análise de dados que são em geral de natureza quantitativa. Há gráficos que suportam a representação de dados qualitativos, como é o caso, por exemplo, do histograma.

Técnicas iconográficas são mais apropriadas para dados quantitativos, pois são os valores dos atributos que fazem variar as propriedades geométricas dos ícones. A representação de dados qualitativos por essas técnicas apresenta mais dificuldades, que podem ser contornadas utilizando propriedades de aparência do ícone, tais como cores.

Técnicas geométricas podem representar dados quantitativos e qualitativos. Já as técnicas orientadas a pixel, como visto nos exemplos na literatura, são empregadas para dados

quantitativos, não sendo recomendadas para dados qualitativos.

Técnicas hierárquicas são ideais para visualizar dados quando estes possuem uma estrutura de relacionamento entre si, seja com uma estrutura de hierarquia ou de simples relação.

Outras classes de dados como documentos, textos (disponíveis na *web* ou armazenados em disco) e códigos de programas são possíveis de serem representados por meio de ferramentas específicas. Para o caso de documentos e textos, trabalhos como os de Havre et al. (2002), Starre e Vries (2005), Kienreich et al. (2005) e Mao et al. (2007), apresentam soluções para a visualização desta classe de dados. A representação visual de códigos de programas, por sua vez, é um assunto discutido em trabalhos como os de Voinea e Telea (2007), Zeckzer et al. (2008) e Sensalire et al. (2008).

6.3. Tipo de tarefa

De modo geral, algumas técnicas servem melhores para determinadas tarefas do que para outras, como é observado na Tabela 6.2. As tarefas estão sujeitas à utilização de técnicas de interação e distorção implementadas pela ferramenta em uso, que melhoram a atividade de exploração dos dados.

Tabela 6.2: Tarefas mais representativas de cada classe de técnica de visualização

Categoria	**Técnicas**	**Tarefas**				
		Visão geral	Correlação	Regras / padrões	Agrupamento	*Outlier*
1-D a 3-D	Histograma			X		X
	Diagrama de caixa			X		X
	Gráfico de dispersão		X	X		X
	Gráfico de contorno		X			
Iconográficas	Faces de Chernoff			X	X	X
	Star Glyphs			X	X	X
	Figuras de Aresta				X	
Geométricas	Matriz de dispersão		X	X		X
	Coordenadas Paralelas	X	X	X		X
Orientadas a pixel	*Query-dependent*		X	X	X	
	Query-independent		X	X	X	
Hierárquicas	Grafos	X	X			
	Cone trees	X				
	Treemap	X	X	X	X	
	Gráfico de mosaico	X		X		
	Empilhamento de dimensões			X	X	X

Seguindo os princípios da Análise Exploratória de Dados (AED), discutida na Seção 2.3.3, as técnicas 1-D a 3-D servem para visualizar uma estimativa de certeza acerca de uma hipótese ou a distribuição de frequência de um atributo, como é o caso da utilização de um histograma. Essa classe também provê gráficos para realizar comparações e classificações de dados e, também, para verificar a correlação entre atributos. A AED incentiva a utilização de diferentes gráficos estatísticos para a análise de dados, levando à descoberta de padrões e estruturas presentes nos dados, bem como identificação de anomalias (*outliers*), que podem ser observadas por meio do diagrama de caixas, por exemplo.

Técnicas iconográficas fornecem meios para a verificação de regras e padrões de comportamentos dos dados, em que cada registro pode ser visualizado utilizando representações gráficas. Ícones com propriedades semelhantes podem ser reconhecidos e, desse modo, formarem grupos que possam ser analisados em particular. Uma representação que apresente um formato discrepante quando comparado às demais pode caracterizar a presença de *outlier*.

Técnicas geométricas fornecem uma boa visão geral dos dados, pois não atribuem prioridades ao representar os atributos e a verificação da correlação entre eles pode ser mais perspicaz utilizando técnicas desta classe como, por exemplo, a matriz de dispersão. Esta classe de técnicas também apóia na identificação de padrões, regras e comportamentos e, por conseguinte, também podem ser analisadas anomalias ou *outliers*, caracterizadas por comportamentos fora do padrão comum. O analista pode optar por analisar um grupo de dados que pode ser destacado por meio da ferramenta em uso, mas, a princípio, agrupamentos não são imediatamente identificados por esta classe de técnicas.

Técnicas orientadas a pixel fornecem meios para análise dos relacionamentos entre os atributos de dados, sendo possível identificar regras e padrões por meio da correlação verificada. Além disso, os pixels podem ser arranjados de forma a indicar agrupamentos.

Técnicas hierárquicas, por sua vez, são úteis para a exploração de dados dispostos em uma estrutura hierárquica ou de relação simples. Por meio dessa classe de técnicas é possível obter uma visão geral da estrutura dos dados, bem como analisar a relação entre os elementos. Técnicas dessa classe também permitem o agrupamento de dados, como é o caso de Treemaps, por exemplo, em que é permitida a visualização de diferentes classificações dos dados, conforme é arranjada a distribuição dos elementos.

6.4. Volume e dimensionalidade

As técnicas de visualização estão relacionadas à dimensionalidade e ao volume dos dados que são capazes de suportar a ponto de proporcionar uma visualização clara ao analista. A Tabela 6.3 traz as proporções mais indicadas de volume e dimensionalidade para cada técnica de visualização analisada neste trabalho.

Tabela 6.3: Técnicas de visualização e respectivas representações de volume e dimensionalidade dos dados

Categoria	Técnicas	Volume	Dimensionalidade
1-D a 3-D	Histograma	Pequeno	Baixa
	Diagrama de caixa		
	Gráfico de dispersão		
	Gráfico de contorno		
Iconográficas	Faces de Chernoff	Pequeno	Baixa a média
	Star Glyphs	Pequeno	
	Figuras de Aresta	Médio	
Geométricas	Matriz de dispersão	Médio	Média a alta
	Coordenadas Paralelas		
Orientadas a pixel	*Query-dependent*	Grande	Média a alta
	Query-independent		
Hierárquicas	Grafos	Pequeno a médio	Alta
	Cone trees	Pequeno a médio	Alta
	Treemap	Médio	Alta
	Gráfico de mosaico	Médio	Média
	Empilhamento de dimensões	Médio	Média

Técnicas de visualização para dados unidimensionais a tridimensionais possuem baixa dimensionalidade e apresentam um volume pequeno de dados, pois em geral são provenientes de cálculos estatísticos, que envolvem uma amostra ou resumem-se em valores percentuais.

Técnicas iconográficas são capazes de lidar com uma quantidade maior de atributos em relação aos gráficos 1D-3D, porém a visualização gerada é melhor para uma quantidade pequena de dados, devido ao espaço ocupado por cada ícone. Essa afirmação é a mesma encontrada em (Rabelo, 2007) em que as técnicas iconográficas avaliadas – *Star Glyphs* e Faces de Chernoff – foram classificadas como técnicas de baixa escalabilidade (suporte a quantidade de dados).

Técnicas geométricas, comparadas às técnicas iconográficas, podem trabalhar com um volume maior de dados, bem como uma quantidade maior de dimensões.

A principal característica apresentada pelas técnicas orientadas a pixel é a possibilidade de representar grande quantidade de dados, já as técnicas hierárquicas e baseadas em grafos preocupam-se em representar o relacionamento entre os dados, não importando a

dimensionalidade, que pode ser baixa ou alta, mas possuem a mesma restrição de espaço que as técnicas iconográficas, em que a visualização é mais clara se a quantidade de dados não é tão volumosa.

Entretanto, as ferramentas de visualização podem oferecer recursos como *zoom*, seleção, destaque, entre outros, para melhorar a interatividade com a visualização, amenizando as limitações de cada técnica.

6.5. Posicionamento dos atributos

Apesar de não ser um parâmetro diretamente ligado às características dos dados, é um fator importante para a tarefa de analisar a correlação entre os atributos. Este parâmetro depende da técnica ou ferramenta utilizada para gerar a visualização, como descrito na Tabela 6.4, que deve possibilitar a mudança das disposições dos atributos, produzindo visualizações diferentes que podem revelar novos padrões.

Devido a isso, na primeira coluna da Tabela 6.4 estão descritas as categorias de técnicas de visualização, seguidas por exemplos de técnicas específicas na segunda coluna e, na terceira, é indicado se o parâmetro de posicionamento do atributo pode influenciar ou não na interpretação da representação gráfica dos dados para a respectiva técnica.

Tabela 6.4: Técnicas de visualização e a interferência do parâmetro posição dos atributos

Categoria	**Técnicas**	**Posição dos atributos**
1D a 3D	Histograma Diagrama de caixa Gráfico de dispersão Gráfico de contorno	Não influencia
Iconográficas	Faces de Chernoff *Star Glyphs* Figuras de Aresta	Não influencia Não influencia Influencia
Geométricas	Matriz de dispersão Coordenadas Paralelas	Não influencia Influencia
Orientadas a pixel	Query-independent Query-dependent technique	Influencia
Grafos	Grafos Cone Trees / Cam Trees *Treemap* Gráfico de mosaico Empilhamento de dimensões	Influencia Não influencia Influencia Influencia Influencia

Técnicas da classe 1D a 3D são aplicados em geral sobre poucos atributos, o que torna o parâmetro posição pouco influente na visualização dos dados. Além disso, o objetivo em usar

esta classe de técnicas determina a alocação dos atributos no gráfico, geralmente dispondo os valores dos dados para um plano cartesiano para: analisar o comportamento de um atributo (no caso de histogramas e diagramas de caixas), verificar a relação entre dois atributos (quando utiliza-se o gráfico de dispersão) ou analisar um atributo em função de outros dois (como no gráfico de contorno).

Das técnicas iconográficas apresentadas na Tabela 6.4, Figuras de Aresta é um exemplo em que a posição dos atributos pode influenciar na exploração visual dos dados conforme são utilizados diferentes formatos de ícones, derivados da variação do mapeamento entre os atributos do dado e as propriedades do ícone (Pickett e Grinstein, 1988; p. 516).

As faces de Chernoff, por sua vez, possuem uma estrutura fixa para o ícone, pois corresponde às características de um rosto humano e, desse modo, a variação das posições dos atributos não é um aspecto relevante para esta técnica. Mas há estudos sobre quais propriedades do ícone podem ser mais representativas para a interpretação dos resultados como, por exemplo, os olhos e o formato da face que são os aspectos que mais chamam a atenção (Morris et al., 2000; Lee et al., 2003). Do mesmo modo é para a técnica *Star Glyph*, para a qual uma vez estabelecida a melhor ordem de mapeamento dos atributos (Peng et al., 2004; Klippel et al. 2009), a posição permanece fixa para todos os ícones que representam um registro por estrela.

Nos trabalhos de Inselberg (2008) e Wegman (1990), é explicado como a posição dos atributos pode influenciar na detecção de correlação nas Coordenadas Paralelas. Já a Matriz de Dispersão, por ser composta de um conjunto de gráficos de dispersão, não sofre grande influência ao variar a posição dos atributos no gráfico, uma vez que o principal objetivo é avaliar a correlação entre os atributos.

Keim (2000) apresenta algoritmos para o posicionamento dos pixels na exibição dos dados, o que pode influenciar a interpretação da visualização quanto à detecção de padrões e relacionamentos entre os atributos representados. A técnica *query-independent* (dirigida a dados), por exemplo, pode ter os pixels arranjados conforme o algoritmo *recursive pattern* (Keim, 2000). A técnica *query-dependent* (dirigida a resultados) pode dispor os pixels na janela seguindo um traço espiral (Keim, 1997).

Técnicas hierárquicas ou baseadas em grafos, em geral, são influenciadas pela posição dos atributos, pois os elementos seguem naturalmente uma estrutura de relacionamento e por isso, deve-se ter cuidado ao alocar os atributos no gráfico, principalmente quando houver uma relação de hierarquia entre os elementos. A exceção é a técnica *Cone Tree*, que representa uma estrutura de árvore já definida (como a estrutura de arquivos e diretórios em um disco rígido), fornecendo apenas funções de interação como a animação para navegar entre os nós da árvore

(Cockburn e McKenzie, 2000; Robertson et al., 1991).

6.6. Considerações finais

Neste capítulo foi realizada uma análise geral sobre a associação entre os parâmetros com as classes de técnicas de visualização. Esta análise resultou nas diretrizes para a escolha de técnicas de visualização. Dessa forma, o objetivo principal proposto por este trabalho foi alcançado, significando, ainda, o encerramento do processo sugerido pela metodologia da Teoria Fundamentada em Dados, sendo a definição dessas diretrizes, após todo o processo de formação da amostragem teórica, codificação e diagramação, equivalente à etapa final de formulação da teoria.

7. Conclusão

As técnicas de visualização podem ser aplicadas durante todo o processo de descoberta de conhecimento. No método tradicional, utilizando algoritmos de mineração de dados, a visualização pode servir para pré-visualizar os dados, visualizar os conhecimentos obtidos ou então representar visualmente os resultados intermediários das iterações do processo de descoberta de conhecimento.

Entretanto, pode-se conseguir mais do poder das técnicas de visualização quando estas são utilizadas diretamente sobre um conjunto de dados dos quais deseja-se obter algum conhecimento novo. Neste contexto, o papel das técnicas de visualização é de uma ferramenta de descoberta de conhecimento pela qual o usuário interage com a visualização, construindo e adquirindo o conhecimento conforme os dados são explorados por meio de sua representação gráfica. Desse modo, o usuário pode optar por utilizar técnicas de visualização para obter informações dos dados sem a necessidade de recorrer a técnicas tradicionais de mineração.

A utilização de técnicas de visualização merece cuidados em relação à sua escolha, que deve levar em consideração alguns fatores que influenciam nesta decisão. Tais fatores foram denominados neste trabalho como parâmetros, que distinguem-se em: tipo de dado, tipo de tarefa, volume, dimensionalidade e posição dos atributos no gráfico.

Esses parâmetros foram analisados teoricamente sobre um conjunto de técnicas de visualização mais frequentemente encontradas na literatura. Posteriormente, foram analisados na prática por meio da utilização de algumas ferramentas computacionais para gerar a representação gráfica de bases de dados.

Estes procedimentos, amparados pela Teoria Fundamentada em Dados, contribuíram para a formulação de diretrizes que indicam quais técnicas de visualização podem ser utilizadas levando em consideração as características dos dados, os objetivos da exploração visual e as funcionalidades presentes nas ferramentas a serem utilizadas.

A Teoria Fundamentada em Dados proveu os métodos necessários para que os

parâmetros pudessem ser identificados e as diretrizes definidas, por meio das etapas de formação da amostragem teórica, codificação, diagramação e formulação da teoria.

A formação da amostragem teórica foi baseada na coleta de dados em fontes bibliográficas e na triangulação, que constituiu-se de resultados da utilização de ferramentas computacionais que implementam técnicas de visualização vistas na literatura pesquisada.

O processo de codificação possibilitou a identificação dos parâmetros para a escolha de técnicas de visualização. Durante a codificação aberta, foi utilizada a estratégia de codificação por pontos-chave, pelo qual surgiram os conceitos que, por meio de comparação constante, foi constatado que repetiam-se nos dados da amostragem teórica. Esses conceitos deram origem aos parâmetros tipo de dado, tipo de tarefa, volume, dimensionalidade e posição dos atributos no gráfico. Na fase de codificação axial, estes parâmetros foram associados e analisados em conjunto com classes de técnica de visualização definidas na taxonomia dada por Keim (2002).

Por meio desta análise, foi possível observar que cada tipo de técnica (gráficos 1D a 3D, iconográficas, geométricas, orientadas a pixel e hierárquicas ou baseadas em grafos) satisfaz a uma determinada configuração desses parâmetros que espelham as características dos dados e variam conforme os objetivos do uso da visualização.

A triangulação foi realizada para demonstrar como a consulta aos parâmetros definidos pode ser feita e também para reafirmar a relevância que esses parâmetros possuem na decisão por técnicas de visualização. Nesse procedimento, foram utilizados preferencialmente softwares com licença de uso gratuita, entre os quais alguns proviam mais de uma técnica de visualização, enquanto que outros, somente uma técnica. Por possuírem implementações particulares, cada ferramenta exige um formato de entrada dos dados diferente. Essa é uma questão que deve ser considerada durante a preparação e seleção dos dados, os quais devem ser formatados para adequarem-se às ferramentas a serem utilizadas durante o processo de extração do conhecimento. Assim, por meio dos processos de codificação e triangulação, foi possível formular as diretrizes.

O tipo de dado deve ser o primeiro parâmetro a ser considerado. É pelo tipo de dado que determina-se qual classe de técnica de visualização pode ser, a priori, utilizada. Dados qualitativos, por exemplo, dificilmente serão inteligíveis quando representados por uma técnica desenvolvida para interpretar dados quantitativos e vice-versa. Além disso, foi verificado por este estudo que há mais opções de técnicas de visualização para representar dados quantitativos do que dados qualitativos.

A tarefa a ser realizada corresponde aos objetivos do analista durante a exploração dos dados. Na literatura são encontradas classificações de técnicas de visualização baseadas neste parâmetro. Para tarefas vinculadas à análise estatística, por exemplo, os gráficos convencionais

que exibem de uma a três dimensões podem ser suficientes; para tarefas de verificação de correlação, podem ser utilizadas técnicas da classe geométrica, e assim por diante.

Os parâmetros volume e dimensionalidade dos dados são fatores limitantes das técnicas de visualização. Apesar da maioria suportar dados multidimensionais, geralmente as técnicas diferem na capacidade de exibir uma quantidade de dimensões e determinado volume de dados. É o caso da diferença entre as classes de técnicas iconográficas, geométricas e orientadas a pixel. Contudo, outras técnicas de interação e distorção podem ser utilizadas durante a exploração visual para amenizar essas limitações como, por exemplo, as funções de *zoom*, seleção e filtro.

O posicionamento dos atributos é um fator mais dependente da técnica de visualização a ser utilizada e, consequentemente, do software que a implementa. Para algumas técnicas, tais como coordenadas paralelas e *star glyphs*, o posicionamento dos atributos é importante para a descoberta de novos padrões ou comportamentos. No caso das coordenadas paralelas, o posicionamento dos atributos influencia no modo em que são exibidas suas linhas poligonais; na técnica *star glyphs*, a ordem de distribuição dos atributos para as propriedades do ícone pode facilitar a tarefa de agrupamento, considerando que diferentes ordens geram diferentes formatos de ícone.

Além dos parâmetros, outro ponto a ser considerado é a familiaridade do analista com os dados analisados. Isto é o que despertará novos interesses ou estimulará a curiosidade do usuário durante a exploração dos dados, formando novas hipóteses que podem ser verificadas por meio das visualizações, ou simplesmente comparando resultados gerados pelas representações gráficas. Diante desse fato, é importante que o analista tenha conhecimento sobre o significado dos dados de modo que a visualização sobre estes possa, de fato, melhorar a interpretação do conhecimento adquirido e estimular a interação do analista com a representação gráfica.

Portanto, a visualização é uma ferramenta vantajosa no entendimento do conhecimento, seja este obtido por meio de algoritmos de mineração de dados, ou pela exploração visual realizada diretamente sobre os dados.

Como sugestão de trabalhos futuros, pode ser citada a análise de outras técnicas de visualização que possam ser aplicadas para as mesmas bases de dados verificadas neste trabalho ou para novos conjuntos de dados, aprimorando a amostragem teórica definida pela TFD.

A principal contribuição deste trabalho foi o estabelecimento das diretrizes para a escolha de técnicas de visualização utilizadas para a extração do conhecimento. Além disso, os parâmetros identificados neste trabalho são critérios importantes para a análise de uma base de dados a ser explorada.

Referências Bibliográficas

ALEXANDRE, D. S.; TAVARES, J. M. R.S. Factores da Percepção Visual Humana na Visualização de Dados. In: *Métodos Numéricos e Computacionais em Engenharia CMNE/CILAMCE*, Porto – Portugal, 2007.

ALLAN, G. A critique of using grounded theory as a research method. *Electronic Journal of Business Research Methods*, v.02, n.03, 2003. Disponível em <http://www.ejbrm.com/>

ANDRIENKO, G.; ANDRIENKO, N. Coordinated Multiple Views: a Critical View. In: *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization* (CMV 07), p. 72-74, Washington, DC, USA, 2007.

ANKERST, M.; BERCHTOLD, S.; KEIM, D. Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data. In: *IEEE Symposium on Information Visualization (InfoVis '98)*, p.52-61, 1998.

ANKERST, M. Visual Data Mining with Pixel-oriented Visualization Techniques. In: *ACM SIGKDD Workshop on Visual Data Mining*, San Francisco, 2001.

ANKERST, M.; ELSEN, C.; ESTER, M.; KRIEGEL, H. Visual Classification: An Interactive Approach to Decision Tree Construction. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, EUA, p. 392-396, 1999.

BAEHRECKE, E. H.; DANG, N.; BABARIA, K.; SHNEIDERMAN, B. Visualization and analysis of microarray and gene ontology data with treemaps. *BMC Bioinformatics*, v. 5, n. 84, 2004.

BALDONADO, M.Q.W.; WOODRUFF, A.; KUCHINSKY, A.. Guidelines for using multiple views in information visualization. In: *Advanced Visual Interfaces*, p.110–119, 2000.

BEDERSON, B. B. PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubblemaps. In: *ACM Symposium on User Interface Software and Technology, CHI Letters*, p. 71-80, 2001.

BEDERSON, B. B.; SHNEIDERMAN, B.; WATTENBERG, M. Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies. *ACM Transactions on Graphics*, v. 21, n. 4, p. 833-854, 2002.

BRUCKNER, L. A. On Chernoff Faces. In: *Symposium on Graphical Representation of Multivariate Data*, California, p. 93-121, 1978.

BRUZZESE, D.; BUONO, P. Combining Visual Techniques for Association Rules Exploration. In: *Proceedings of the working conference on Advanced Visual Interfaces* – ACM, Itália, p. 381-384, 2004.

BURKHARD, R. A. Impulse: Using Knowledge Visualization in Business Process Oriented Knowledge Infrastructures. *Journal of Universal Knowledge Management*, n. 2, p.170-188, 2005.

BURKHARD, R. A. Learning from Architects: The Difference between Knowledge Visualization and Information Visualization. In: *Proceedings of the Eighth International Conference on Information Visualisation (IV'04)*, IEEE Computer Society, London, England, p. 519-524, 2004.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. *Readings in Information Visualization - Using Vision to Think*, San Francisco: Ed.Morgan Kaufmann, 1999.

CARR, D. B.; LITTLEFIELD, R. J.; NICHOLSON, W. L.; LITTLEFIELD, J. S. Scatterplot Matrix Techniques for Large N. *Journal of the American Statistical Association*, v. 82, n. 398, p. 424-436, 1987.

CHEN, C.; HÄRDLE, W.; UNWIN, A. *Handbook of Data Visualization*. Ed. Springer, 2008.

CHERNOFF, H. The Use of Faces to Represent Points in K-Dimensional Space Graphically. *Journal of the American Statistical Association*, v.68, n. 342, p.361-368, 1973.

CHI, E. H. A Taxonomy of Visualization Techniques using the Data State Reference Model. In: *Proceedings of the IEEE Symposium on Information Visualization*, 2000.

COCKBURN, A.; MCKENZIE, B. An Evaluation of Cone Trees. In: *People and Computers XIV: British Computer Society Conference on Human Computer Interaction*, p. 425-436, 2000.

COLEMAN, G.; O'CONNOR, R. Using grounded theory to understand software process improvement: A study of Irish software product companies. *Elsevier – Information and Software Technology*, v. 49, p. 654–667, 2007.

COULOM, R. Treemaps for Search-Tree Visualization. In: *The 7th Computer Olympiad Computer-Games Workshop*, p. 1-7, 2002.

DANTAS, C. C.; LEITE, J. L.; LIMA, S. B. S.; STIPP, M. A. C. Teoria Fundamentada nos Dados - Aspectos Conceituais E Operacionais: Metodologia Possível de Ser Aplicada na Pesquisa em Enfermagem. *Revista Latino-americana de Enfermagem*, v.17, n.4, p. 573-579, 2009.

DAVENPORT, T.H.; PRUSAK, L. *Working Knowledge: How Organizations Manage What They Know*. Boston, MA:Harvard Business School Press, 1998.

DIAS, M. M. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. *Acta Scientiarum. Technology*, Ed. EDUEM, v. 24, n.6, p. 1715-1725, 2002.

DICK, B. Grounded theory: A thumbnail sketch. Disponível em <http://www.scu.edu.au/schools/gcm/ar/arp/grounded.html>, 2005. Acesso em 3 de janeiro de 2010.

EICK, S. G.; WILLS, G. J. High interaction graphics. *European Journal of Operational Research. Elsevier Science B.V.*, v. 81, n. 3, p. 445-459, 1995.

EPPLER, M.J.; BURKHARD, R.A. Knowledge Visualization. Towards a new discipline and its field of application, 2004. Disponível em <http://www.bul.unisi.ch/cerca/bul/pubblicazioni/com/pdf/wpca0402.pdf>. Acesso em 29 de abril de 2009.

ESTEVES, J.; RAMOS, I.; CARVALHO, J. A. Use of Grounded Theory in Information Systems Area : An Exploratory Analysis. In: *Proceedings of European Conference on Research Methods*

(ECRM), MCIL, ISBN: 0-9540488-3-0, p. 129-136, 2002.

EVL – Electronic Visualization Laboratory. Disponível em <http://www.evl.uic.edu/aej/526/kyoung/Training-parallelcoordinate.html> Acessado em 16 de nov. de 2009.

FANEA, E.; CARPENDALE, S.; ISENBERG, T. An Interactive 3D Integration of Parallel Coordinates and StarGlyphs. In: *Proceedings of the 2005 IEEE Symposium on Information Visualization*, 2005.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, v. 39, n. 11, p. 27-34, 1996.

FELDENS, M. A; MORAES, R. L.; PAVAN, A.; CASTILHO, J. M. V. Towards a methodology for the discovery of useful knowledge combining data mining, data warehousing and visualization. In: *CLEI – Conferência Latino-Americana de Informática*, Quito, Equador, v. 24, 1998.

FILIPPO, D. D. R. *Suporte à coordenação de sistemas colaborativos: uma pesquisa-ação com aprendizes e mediadores atuando em fóruns de discussão de um curso a distância*. Tese de doutorado, Pontifícia Universidade Católica do Rio de Janeiro, PUC-Rio, Rio de Janeiro, 2008.

FLURY, B.; RIEDWYL, H. Graphical Representation of Multivariate Data by Means of Asymmetrical Faces. *Journal of the American Statistical Association*, v. 76, n. 376, p. 757-765, 1981.

FREITAS, C. M. D. S.; CHUBACHI, O. M.; LUZZARDI, P. R. G.; CAVA, R. A. Introdução à Visualização de Informações. *Revista de Informática Teórica e Aplicada - RITA*, vol. VIII, n. 02, 2001.

FREITAS, D.; JACKIX, M.. Caracterização Físico-Química e Aceitação Sensorial de Bebida Funcional Adicionada de Frutoligos-Sacarídeo e Fibra Solúvel. *Boletim do Centro de Pesquisa de Processamento de Alimentos*, América do Sul, 2004.

FRIENDLY, M. A Brief History of Data Visualization. In: CHEN, C.; HÄRDLE, W.; UNWIN, A. *Handbook of Data Visualization*, Springer, 2008, p. 15-56.

FRIENDLY, M. Extending Mosaic Displays: Marginal, Partial, and Conditional Views of Categorical Data. *Data Visualization in Statistics Conference*, 1998.

FRIENDLY, M; DENIS, D. Marcos na História da Visualização de Dados – Traduzido e organizado por KANNO, M. Disponível em <http://www.math.yorku.ca/SCS/Gallery/milestone/>. Acesso em 25 mar. 2009.

GANESH, M.; HAN, E.; KUMAR, V.; SHEKHAR, S.; SRIVASTAVA, J. Visual Data Mining: Framework and Algorithm Development, 1996 Disponível em <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.2401>. Acesso em 20 de agosto de 2009.

GONICK, L.; SMITH, W. The Cartoon Guide to Statistics, Harper Perennial, p. 212, 1993. Extraído de Wolfram MathWorld, Disponível em <http://mathworld.wolfram.com/ChernoffFace.html>. Acessado em 06 de novembro de 2009.

GREEN, M. Toward a Perceptual Science of Multidimensional Data Visualization: Bertin and Beyond. *ERGO/GERO Human Factors Science*, 1998. Disponível em <http://graphics.stanford.edu/courses/cs448b-06-winter/papers/Green_Towards.pdf>. Acessado em 06 de novembro de 2009.

GRINSTEIN, G.; TRUTSCHL, M.; CVEK, U. High-Dimensional Visualization. In: *Proceedings of the Visual Data MiningWorkshop (KDD'01)*, 2001.

HAIG, B. D. Grounded Theory as Scientific Method. Philosophy of Education, 1995. Disponível em < http://www.ed.uiuc.edu/EPS/PES-Yearbook/95_docs/haig.html>. Acesso em 03 de fev. de 2010.

HAN, J.; KAMBER, M. *Data mining: concepts and techniques*. USA: The Morgan Kaufmann, 2001.

HAVRE, S.; HETZLER, E.; WHITNEY, P.; NOWELL, L. ThemeRiver: Visualizing Thematic Changes in Large Document Collections. *IEEE Transactions on Visualization and Computer Graphics*, v. 08, n.1, p. 9-20, 2002.

HERMAN, I.; MELANÇON, G.; MARSHALL, S. Graph Visualization and Navigation in Information Visualization: A Survey. *IEEE Transactions on Visualization and Computer Graphics*, v. 06, n.1, p. 24-43, 2000.

HINNEBURG, A.; KEIM, D.; WAWRYNIUK, M. HD-Eye: Visual Mining of High-Dimensional Data. *IEEE Computer Graphics and Applications*, v. 19, n. 5, p. 22-31, 1999.

HOFFMAN, P. E. *Table Visualizations: A Formal Model and Its Applications*. Tese de doutorado, University of Massachusetts Lowell, 1999.

HOFMANN, H.; SIEBES, A. P. J. M.; WILHELM, A. F. X. Visualizing Association Rules with Interactive Mosaic Plots. In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, EUA, p. 227-235, 2000.

HOFMANN, H. Mosaic Plots and Their Variants. In: CHEN, C.; HÄRDLE, W.; UNWIN, A. *Handbook of Data Visualization*, Springer, 2008, p.618 - 642.

HOFMANN, H. Constructing and reading mosaicplots. *Elsevier – Computational Statistics & Data Analysis*, v. 43, p. 565–580, 2003.

HUNTER, K.; HARI, S.; EGBU, C.; KELLY, J. Grounded Theory: Its Diversification and Application Through two Examples From Research Studies on Knowledge and Value Management. *The Electronic Journal of Business Research Methodology*, v. 3, n. 1, p. 57-68, 2005. Disponível em <www.ejbrm.com>.

INMON, W. H. *Como Construir o Data Warehouse*. 2 ed. Rio de Janeiro: Campus, 1997.

INSELBERG, A. The plane with parallel coordinates. *The visual Computer, Springer*, v.1, n.2, , p. 69-91, 1985.

INSELBERG, A., DIMSDALE, B. Parallel coordinates: a tool for visualizing multidimensional geometry. In: *Proceedings of Visualization '90*, p. 361 - 378, 1990.

INSELBERG, A. Parallel Coordinates: Visualization, Exploration and Classification of High-Dimensional Data. In: CHEN, C.; HÄRDLE, W.; UNWIN, A. *Handbook of Data Visualization*, Springer, 2008, p. 643-680.

JOHNSON, B.; SHNEIDERMAN, B. Tree-Maps: a space-filling approach to the visualization of hierarchical information structures. In: *Proceedings of the 2nd Conference on Visualization '91*, San Diego, California, IEEE Computer Society Press, p. 284-291, 1991.

KEIM, D. A.; KRIEGEL, H. Visualization Techniques for Mining Large Databases: A Comparison. *IEEE Transactions on knowledge and data engineering*, v. 8, n. 6, p. 923-938, 1996.

KEIM, D. A. Pixel-oriented Visualization Techniques for Exploring Very Large Databases. *Journal of Computational and Graphical Statistics*, v.5, p. 58-77, 1996.

KEIM, D. A. Visual Techniques for Exploring Databases. Invited Tutorial, *Int. Conference on Knowledge Discovery in Databases (KDD'97)*, Newport Beach, CA, 1997.

KEIM, D. Designing Pixel-Oriented Visualization Techiniques: Theory and Applications. *IEEE Transactions on Visualization and Computer Graphics*, v. 6, n.1, p. 1-20, 2000.

KEIM, D. A. Visual Exploration of Large Datasets. *Communications of the ACM*, v. 44, n. 08, p. 39 – 44, 2001.

KEIM, D. A. Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, v. 7, n. 1, p. 100-107, 2002.

KEIM, D. A.; SCHNEIDEWIND, J.; SIPS, M. CircleView: a new approach for visualizing time-related multidimensional data sets. In: *Proceedings of the Working Conference on Advanced Visual interfaces (AVI '04)*, Gallipoli, Itália, 2004.

KEIM, D. A. Scaling Visual Analytics to Very Large Data Sets. In: *Workshop on Visual Analytics*, Darmstadt, 2005.

KEIM, D. A.; ANDRIENKO, G.; FEKETE, J. D. ; GÖRG, C., KOHLHAMMER, J.; MELAÇON, G. Visual Analytics: Definition, Process, and Challenges. In: KERREN, A. et al. *Information Visualization-Human-Centered Issues and Perspectives*, Springer, 2008, p. 161–182.

KIENREICH, W.; SABOL, V.; GRANITZER M.; KLIEBER, W.; LUX, M.; SARKA, W. A Visual Query Interface for a Very Large Newspaper Article Repository. In: *Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05), IEEE Computer Society*, 2005.

KITCHENHAM, B.; PICKARD, L.; PFLEEGER, S. L. Case Studies for Method and Tool Evaluation. *IEEE Software*, p. 52-62, 1995.

KLIPPEL, A.; HARDISTY, F.; WEAVER, C. Star Plots: How Shapes Characteristics Influence Classification Tasks. *Cartography and Geographic Information Science*, 2009.

KOCK JR., N.F.; MCQUEEN, R.J.; BARKER, M.Learning and Process Improvement in Knowledge Organisations: A Critical Analysis of Four Contemporary Myths. *The Learning Organization*, p. 31–40, 1996.

KOPANAKIS, I.; THEODOULIDIS, B. Visual data mining modeling techniques for the visualization of mining outcomes. *Elsevier – Journal of Visual Languages and Computing*, 2003.

LARAMEE, R. S. HAUSER, H.; DOLEISCH, H.; VROLIJK, B.; POST, F.; WEISKOPF, D. The State of the Art in Flow Visualization: Dense and Texture-Based Techniques. *Computer Graphics Forum*, v. 23, n. 2., p. 203-221, 2004.

LEBLANC, J.; WARD, M. O.; WITTELS, N. Exploring N-dimensional Databases. In: *IEEE Computer Society Technical Committee on Computer Graphics*, p.230-237, 1990.

LEE, M. D.; REILLY, R. E.; BUTAVICIUS, M. E. An empirical evaluation of Chernoff faces, star glyphs, and spatial visualizations for binary data. In: *Proceedings of the Asia-Pacific Symposium on information Visualisation*, Australia, v. 24, 2003.

LUGLI. Disponível em <http://www.lugli.org/2008/02/diagrama-de-dispersao/>. Acesso em 17 de novembro de 2009.

MATAVIRE, R.; BROWN, I. Investigating the use of "Grounded Theory" in information systems research. In: *SAICSIT '08: Proceedings of the 2008 annual research conference of the South African Institute of Computer Scientists and Information Technologists on IT research in developing countries*, Africa do Sul, p. 139-147, 2008.

MAO, Y.; DILLON, J. V.; LEBANON, G. Sequential Document Visualization. *IEEE Transactions on Visualization and Computer Graphics*, v. 13, n.6, p. 1208-1215, 2007.

MENDONÇA, M.; SUNDERHAFT, N. Mining Software Engineering Data: A Survey State-of-the-Art Report (SOAR). Relatório técnico, *DACS (Data & Analysis Center for Software)*, p. 6-168, 1999.

MYATT, G. J. *Making Sense of Data - A Practical Guide to Exploratory Data Analysis and Data Mining*, John Wiley & Sons, Inc. 2007.

MORGENTHALER, S. Exploratory Data Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, v.1, n.1, p. 33 – 44, 2009. Disponível em <http://www3.interscience.wiley.com/journal/122511655/abstract>. Acesso em 24 de julho de 2009.

MORRIS, C. J.; EBERT, D. S.; RHEINGANS, P. An Experimental Analysis of the Effectiveness of Features in Chernoff Faces. In: *Proceedings of SPIE, 28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making*, *Proceedings of SPIE*, v. 3905, p.12–17, 2000.

NASCIMENTO, H. A. D.; FERREIRA, Cristiane B. R. Visualização de Informações – Uma Abordagem Prática. In: *Anais do XXV Congresso da Sociedade Brasileira da Computação*. Unisinos, São Leopoldo – RS, p. 1262 – 1312, 2005.

NETO, M. A. S. *Mineração Visual de Dados: extração do conhecimento a partir das técnicas de visualização de informação e mineração de dados – Experimentos: Itaipu e Simepar*. Dissertação de mestrado, Universidade Federal do Paraná, Curitiba, 2008.

NIST/SEMATECH – *e-Handbook of Statistical Methods*. Disponível em <http://www.itl.nist.gov/div898/handbook/>. Data da última atualização em 18 de julho de 2006. Acesso em 20 de julho de 2009.

OLIVEIRA, M. C. F.; LEVKOWITZ, H. From Visual Data Exploration to Visual Data Mining: A Survey. In: *IEEE Transactions on Visualization and Computer Graphics*, v.9, n. 3, p. 378-394, 2003.

OLIVEIRA, C. S.; SOUZA, C. R. B.; REIS, C. A. L. Estudo da Alocação de Pessoas em Projetos de Software através da Teoria Fundamentada em Dados. In: *VI Experimental Software Engineering Latin American Workshop, ESELAW'09*, p. 32-41, 2009.

ORLIKOWSKI, W. J. CASE Tools as Organizational Change:Investigating Incremental and Radical Changes in Systems Development. *Management Information Systems Quarterly*, v. 17, n. 3, 1993.

PENG, W.; WARD, M. O.; RUNDENSTEINER, E. A. Clutter reduction in multidimensional data visualization using dimensional reordering. In: *Proceedings of the IEEE Symposium on Information Visualization*, p.89–96, 2004.

PICKETT, R. M.; GRINSTEIN, G. G. Iconographic Displays for Visualizing Multidimensional Data. In: *Proc. IEEE Conference on Systems, Man and Cybernetics, IEEE Press*, Piscataway, NJ, p.514-519, 1988.

PILLAT, R. M.; VALIATI, E. R. A.; FREITAS, C. M. D. S. Experimental Study on Evaluation of Multidimensional Information Visualization Techniques. In: *Proceedings of the 2005 Latin American conference on Human-computer interaction*, p. 20-30, 2005.

RABELO, E. *Avaliação de Técnicas de Visualização de Informação para Mineração de Dados.* Dissertação de mestrado, Universidade Estadual de Maringá, Maringá, 2007.

RABELO, E.; DIAS, M. M.; FRANCO, C.; PACHECO, R. C. S. Information Visualization Which the most Appropriate Technique to Represent Data Mining Results? In: *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation*, Viena, Austria, p. 1218-1223, 2008.

REZENDE, S. O. Mineração de Dados. In: *XXV Congresso da Sociedade Brasileira de Computação*, Unisinos, São Leopoldo, RS, p.397-433, 2005.

RHYNE, T. M. Visualization Viewpoints - Does the Difference between Information and Scientific Visualization Really Matter? *IEEE Computer Graphics and Applications*, v.23, n.03, p. 6-8, 2003.

ROBERTS, J. C. State of the Art: Coordinated & Multiple Views in Exploratory Visualization. In: *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, Washington, DC, USA, p. 61-71, 2007.

ROBERTSON, G. G.; MACKINLAY, J. D.; CARD, S. T. Cone Trees: Animated 3D visualizations of hierarquical information. In: *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, p. 189-194, 1991.

RODRIGUES, A.; ALBUQUERQUE, C. H. L.; BENTO, C. A. C.; VIEIRA, J. M. A.; SILVA, J. G. A. O. B. *Grounded Theory: Problemas de "Alicerçagem"*. Relatório Técnico. Universidade de Coimbra, Portugal, 2004.

RODON, J.; PASTOR, J. A. Applying Grounded Theory to Study the Implementation of an Inter-Organizational Information System**.** *Electronic Journal of Business Research Methods*, v.5, n.2, 2007. Disponível em <http://www.ejbrm.com>

SAXENA, P. C.; NAVANEETHAM, K. The Effect of Cluster Size, Dimensionality, and Number of Clusters on Recovery of True Cluster Structure Through Chernoff-Type Faces. *Journal of the Royal Statistical Society*, v. 40, n. 4, p. 415-425, 1991.

SEIDEL, S. Y.; RAPPAPORT, T. S. 914 Mhz Path Loss Prediction Models for Indoor Wireless Communications in Multifloored Buildings. *IEEE Transactions on Antennas and Propagation*, v. 40, n. 02, 1992.

SENSARILE, M.; OGAO, P.; TELEA, A. Classifying desirable features of software visualization tools for corrective maintenance. In: *Proceedings of the 4th ACM symposium on Software visualization*, Ammersee, Germany, p.87-90, 2008.

SHIMABUKURO, Milton H. *Visualizações Temporais em uma Plataforma de Software Extensível e Adaptável*, Tese de doutorado, Universidade de São Paulo – São Carlos, 2004.

SCHOLTZ, J. Beyond Usability: Evaluation Aspects of Visual Analytic Environments. In: *IEEE Symposium on Visual Analytics Science and Technology*, Baltimore, USA, p. 145-150, 2006.

SHNEIDERMAN, B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: *Proceedings of IEEE Symposium on Visual Languages*, Boulder, CO, p. 336-343, 1996.

SHNEIDERMAN, B. *Discovering Business Intelligence Using Treemap Visualizations.* Relatório técnico, Human Computer Interaction Lab – HCIL, University of Maryland, 2006.

SILVA, E. L.; MENEZES, E. M. *Metodologia da pesquisa e elaboração de dissertação*. 4. ed. Florianópolis: Universidade Federal de Santa Catarina – UFSC, 2005. 138 p. Disponível em: <http://www.portaldeconhecimentos.org.br/index.php/por/content/view/full/10232>. Acesso em: 11 maio 2009.

SILVA, S. L. Gestão do conhecimento: uma revisão crítica orientada pela abordagem da criação do conhecimento. *Ci. Inf., Brasília*, v. 33, n. 2, p. 143-151, 2004.

SIMONI, C. A. C.; BARANAUSKAS, M. C. C. Pesquisa Qualitativa em Sistemas de Informação. Relatório Técnico, Universidade Estadual de Campinas – Unicamp, Instituto de Computação, 63 p., 2003.

STARRE, L.; VRIES, T. Visualizing documents: analysis and evaluation. 2005. Disponível em <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.128.5023>. Acesso em: 26 de junho de 2009.

TAYLOR, A. L.; HICKEY, T. J.; PRINZ, A. A.; MARDER, E. Structure and Visualization of High-Dimensional Conductance Spaces. *Journal of Neurophysiology*, v. 96, p. 891-905, 2006.

TELES, C. D.; FLÔRES, S. H. Influência da adição de espessantes e leite em pó nas características reológicas do iogurte desnatado. *Boletim do Centro de Pesquisa e Processamento de Alimentos*, v. 25, n.02, 2007.

THEUS, M. Interactive data visualization using Mondrian. *Statistical Computing & Statistical Graphics Newsletter*, 2002.

THEUS, M. High-dimensional Data Visualization. In: CHEN, C.; HÄRDLE, W.; UNWIN, A. *Handbook of Data Visualization*, Springer, 2008, p. 151- 178.

THOMAS, J.; COOK, K. Visualization viewpoints: a visual analytics agenda. *IEEE Computer Graphics And Applications*, v.8, n. 2, p. 10-13, 2006.

TUOMI, I. Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organization memory. *Journal of Management Information Systems*, v. 16, n. 3, p. 103-117, Winter 1999.

UNWIN, A. Good Graphics? In: CHEN, C.; HÄRDLE, W.; UNWIN, A. *Handbook of Data Visualization*, Springer, 2008, p. 57-78.

URBANEK, S. Visualizing Trees and Forests. In: CHEN, C.; HÄRDLE, W.; UNWIN, A. *Handbook of Data Visualization*, Springer, 2008, p. 243-264.

VALIATI, E. R. A. *Avaliação de usabilidade de técnicas de visualização de informações multidimensionais*. Tese de doutorado, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2008.

VOINEA, L.; TELEA, A. Visual data mining and analysis of software repositories. *Elsevier – Computer & Graphics*, v. 31, p. 410-428, 2007.

WAINER, J. Métodos de pesquisa quantitativa e qualitativa para a ciência de computação. In: *Jornada de Atualização em Informática, Anais do XXVII Congresso da Sociedade Brasileira de Computação* – JAI, 2007.

WALKER, D.; MYRICK, F. Grounded Theory: An Exploration of Process and Procedure. *Qualitative Health Research*, v. 16, n. 4, p. 547-559, 2006.

WALSUM, T.; POST, F. H.; SILVER, D.; POST, F. J. Feature Extraction and Iconic Visualization. *IEEE Transactions on Visualization and Computer Graphics*, p. 111-119, 1996.

WARD, M. O. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, v. 1, n. 3, p. 194-210, 2002.

WARD, M. O.; LEBLANC, J. T.; TIPNIS, R. N-Land: a Graphical Tool for Exploring N-Dimensional Data. In: *Computer Graphics International Conference in Melbourne*, Australia, 1994.

WARD, M. O. Multivariate Data Glyphs: Principles and Practice. In: CHEN, C.; HÄRDLE, W.; UNWIN, A. *Handbook of Data Visualization*, Springer, 2008, p. 179-198.

WEGMAN, E. J. Hyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statistical Association*, v. 85, n. 411, p. 664-675, 1990.

WILKINSON, L. Graph-theoreticGraphics. In: CHEN, C.; HÄRDLE, W.; UNWIN, A. *Handbook of Data Visualization*, Springer, 2008, p. 121- 150.

WILLS, G. Linked Data Views. In: CHEN, C.; HÄRDLE, W.; UNWIN, A. *Handbook of Data Visualization*, Springer, 2008, p. 217-241.

WILHELM, A. Linked Views for Visual Exploration. In: CHEN, C.; HÄRDLE, W.; UNWIN, A. *Handbook of Data Visualization*, Springer, 2008, p. 199-215.

WYATT, R**.** Face charts: a better method for visualizing complicated data. In: *Proceedings of the IADIS International Conference on Computer Graphics and Visualization*, Amsterdam, p. 51-59, 2008.

ZECKZER, D.; KALCKLÖSCH, R.; SCHRÖDER, L.; HAGEN, H.; KLEIN, T. Analyzing the reliability of communication between software entities using a 3D visualization of clustered graphs. In: *Proceedings of the 4th ACM symposium on Software visualization*, Alemanha, p. 37-46, 2008.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo