FERNANDO ALMIR NASCIMENTO JÚNIOR

VISUALIZAÇÃO DE REGRAS DE ASSOCIAÇÃO

Belo Horizonte

30 de maio de 2005

FERNANDO ALMIR NASCIMENTO JÚNIOR

VISUALIZAÇÃO DE REGRAS DE ASSOCIAÇÃO

Dissertação apresentada ao Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Belo Horizonte

30 de maio de 2005

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Visualização de Regras de Associação

FERNANDO ALMIR NASCIMENTO JÚNIOR

Dissertação defendida e aprovada pela banca examinadora constituída por:

Prof. Wagner Meira Júnior – Orientador
Universidade Federal de Minas Gerais

Profa. Raquel Oliveira Prates
Universidade Estadual do Rio de Janeiro

Prof. Clarindo Isaías Pereira da Silva e Pádua
Universidade Federal de Minas Gerais

Prof. Dorgival Olavo Guedes Neto
Universidade Federal de Minas Gerais

Prof. Renato Antônio Celso Ferreira
Universidade Federal de Minas Gerais

Belo Horizonte, 30 de maio de 2005

Resumo

A mineração de dados é uma área de pesquisa que vem recebendo muita atenção nos últimos anos, pelo seu enorme potencial de extrair informações úteis de grandes volumes de dados. Entretanto, a despeito da sua grande aplicabilidade, as técnicas de mineração de dados ainda não ganharam ampla aceitação entre os usuários leigos, em função principalmente da complexidade dos conceitos envolvidos, que não fazem parte do domínio desses usuários. Uma das mais populares tarefas de mineração de dados é a mineração de regras de associação, que tem aplicação em diferentes contextos, como detecção de fraude em compras públicas, ciências sociais e marketing. A mineração de regras de associação apresenta dois grandes problemas relacionados à interação humana. Em primeiro lugar, o volume de regras geradas é muito grande, o que dificulta a identificação das regras mais interessantes. Em segundo lugar, os conceitos relacionados à técnica são complexos, e quando o usuário não consegue compreender esses conceitos, ele não consegue utilizar o sistema de forma satisfatória.

Nesta dissertação, abordamos esses dois problemas. Para o problema de identificação das regras mais interessantes, apresentamos duas estratégias de visualização do conjunto de regras geradas. Além disso, mostramos como a segunda estratégia supera a primeira, com base em avaliações com usuários e na literatura. Para o segundo problema, identificamos, a partir da literatura e da nossa experiência com o sistema Tamanduá, os principais aspectos que devem ser comunicados pelos projetistas de sistemas de mineração de regras de associação aos usuários, de acordo com a teoria da engenharia semiótica. Para cada um dos aspectos levantados, discutimos sua importância para a interação e o custo para o usuário de ele não ser bem comunicado através da interface.

Abstract

Data mining has been receiving significant attention as a research area in the last decade, as a consequence of its huge potential for extracting useful information from large data volumes. However, despite its applicability, data mining systems did not get wide acceptance among users who are not data mining experts, mainly because of the complexity of its concepts, which are not in the user's knowledge domain. One of the most popular data mining tasks is mining association rules, which has applications in several scenarios such as fraud detection in public expenses, social sciences, and marketing. Mining association rules poses two big challenges related to human-computer interaction. First, the volume of rules generated is very large, being hard to determine the most interesting ones. Second, the premises of the techniques are usually complex, and when the user does not understand these concepts, he is not able to exploit the system resources completely.

In this thesis, we address these two challenges. We propose two visualization strategies of the mined rules, which help identifying the most interesting ones. Further, we discuss how the second strategy improved the user experience, based on user evaluations and other evaluation works. For the second problem, we identify, based on related works and our own experience observing Tamanduá users, the main aspects related to association-rules data mining systems that should be provided to users by system designers, according to semiotic engineering principles. For each aspect evaluated, we discuss its importance in terms of the interaction quality and the cost of not being aware of the interface and its functionalities.

Agradecimentos

Ao professor Wagner Meira, pelo entusiasmo, dedicação e paciência, pelos ensinamentos inestimáveis e principalmente pelo interesse genuíno que nutre pelo crescimento intelectual dos seus alunos.

Aos amigos que fiz ao longo do curso e na convivência no laboratório, em especial aqueles com quem trabalhei mais diretamente: Elisa, Leo Rocha, Bruno Grossi, Juliano, Luiz G., Emílio, Thiago, Yuri, Macambira e Coutinho.

Ao amigo Wagner Toledo, pelo apoio e inspiração, principalmente nas primeiras etapas deste trabalho, e por ter me apresentado a área de Visualização da Informação.

À professora Raquel Prates, pela dedicação e ensinamentos, principalmente nas etapas finais deste trabalho, e por ter me apresentado a Engenharia Semiótica.

Aos professores Dorgival e Renato, pelo suporte dado ao longo de todo o desenvolvimento deste trabalho.

Ao professor Clarindo, pelo que me ensinou sobre Engenharia de Usabilidade e pelas dicas valiosas sobre como melhorar o meu trabalho.

Aos professores Christiano Becker e Loureiro, por acreditarem no meu potencial.

Aos amigos que acompanharam de fora o meu trabalho, torceram por mim e me incentivaram, e principalmente demonstraram muita paciência.

E sobretudo aos meus pais, Fernando e Iara, à minha irmã, Grazielly, ao meu sobrinho, Matheus, e àquela que considero minha segunda família, Luiz Carlos, Tia Marília, Luiz Felipe e Mariana.

Sumário

Lista de Figuras

Capítulo 1

Introdução

O incrível aumento do poder computacional, disponível a preços cada vez mais acessíveis, associado a uma crescente automatização de tarefas, tem levado ao armazenamento de volumes cada vez maiores de dados nas organizações e centros de pesquisa. A mineração de dados surgiu recentemente como uma alternativa promissora para a análise desses grandes volumes de dados. Conjugando técnicas provenientes de diversas áreas, como estatística e banco de dados, a mineração de dados se diferencia das demais técnicas de análise pelo seu caráter exploratório. Se na estatística prevalecem os testes de hipótese e em bancos de dados as consultas estruturadas, na mineração de dados prevalece a detecção automática de padrões. Ou seja, sem que se necessite formular previamente nenhuma hipótese, toda a base de dados é analisada e uma série de padrões explicitados, fornecendo ao analista um conjunto de hipóteses potenciais que, dado o tamanho da base, só poderiam ser levantadas através da intuição. Em poucas palavras, a mineração de dados pode ser definida como a "extração de informações implícitas, previamente desconhecidas e potencialmente úteis de grandes bases de dados" (Witten e Frank, 2000).

Uma tarefa de mineração de dados envolve várias etapas (Figura 1.1). A primeira etapa consiste na compreensão do problema que se deseja resolver, do ponto de vista do domínio de aplicação, e no mapeamento deste problema em um problema de mineração de dados. A segunda etapa compreende a seleção e a preparação dos dados a serem minerados. Na terceira etapa, acontece a mineração propriamente dita, quando os padrões são descobertos e explicitados. A quarta e última etapa consiste na visualização dos resultados – ou modelos – e na obtenção do conhecimento pelo usuário. Ao conjunto dessas etapas dá-se o nome de *descoberta de conhecimento em bancos de dados* (Fayyad et al., 1996). Entretanto, o termo mineração de dados, que antes era usado para referenciar apenas a terceira etapa do processo, se popularizou como sinônimo do processo como um todo.

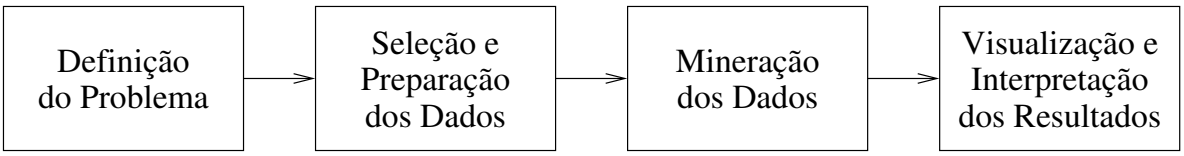


Figura 1.1: Descoberta de conhecimento em bancos de dados.

Dentre as tarefas de mineração de dados, três merecem destaque especial: análise de agrupamentos, classificação e mineração de regras de associação. A **análise de agrupamentos** é usada sempre que se deseja segmentar uma base de dados de acordo com algum critério de similaridade. Por exemplo, se uma empresa quiser segmentar os seus clientes para identificar os perfis mais significativos, e desta forma poder direcionar melhor as suas ações de marketing, a tarefa de mineração de dados escolhida deverá ser a análise de agrupamentos. Os padrões minerados em uma tarefa deste tipo são chamados de grupos ou agrupamentos (do inglês *clusters*). No exemplo citado, são minerados grupos de clientes, sendo que os clientes dentro de um mesmo grupo apresentam perfis semelhantes e clientes em grupos diferentes apresentam perfis distintos.

A **classificação** é usada sempre que se deseja prever o valor de um determinado atributo para um dado registro em uma base de dados. O atributo cujo valor se deseja prever é chamado de classe-alvo. Suponhamos, por exemplo, que uma empresa deseja prever quais dos seus atuais clientes têm mais chance de deixarem de ser seus clientes para se tornarem clientes de uma empresa concorrente. Este fenômeno é conhecido em marketing como *churn* e causa enormes prejuízos às empresas. A classe-alvo neste caso é um atributo que diz se um cliente é um cliente fiel ou um ex-cliente. Uma tarefa de classificação permitiria que a empresa identificasse as características que melhor distinguem um cliente fiel de um ex-cliente, possibilitando a ela prever eminentes perdas de clientes.

A terceira tarefa de mineração de dados é a chamada **mineração de regras de associação**. Como este trabalho é sobre a visualização de regras de associação, dedicamos a próxima seção à explicação dos conceitos que envolvem esta tarefa de mineração de dados.

1.1 Mineração de Regras de Associação

A mineração de regras de associação é uma das mais populares tarefas de mineração de dados, tendo sido introduzida por Agrawal et al. (1993). A aplicação canônica da mineração de regras de associação é a chamada *análise do carrinho de compras*, que consiste na compreensão dos hábitos de compra dos clientes de um supermercado. A

idéia é descobrir como as vendas de alguns produtos influenciam nas vendas de outros produtos, para que se possa planejar melhor as promoções, organizar de forma mais conveniente a disposição das prateleiras e avaliar o impacto que a descontinuidade nas vendas de um produto pode provocar nas vendas de outros produtos.

Uma regra de associação representa então uma relação entre dois ou mais itens de uma base de dados. Considere por exemplo a seguinte regra, que poderia ter sido minerada na base de transações de uma padaria hipotética:

$$[\text{Pão}], [\text{Manteiga}] \Rightarrow [\text{Leite}] \ (80.00, 50.00)$$

Esta regra mostra a relação que existe entre a compra de Pão, Manteiga e Leite nessa padaria e deve ser lida da seguinte forma: cinqüenta por cento das compras realizadas pelos clientes da padaria incluem Pão, Leite e Manteiga; e das compras que incluem Pão e Manteiga, oitenta por cento também incluem Leite. O conjunto de itens do lado esquerdo da regra (Pão e Manteiga) é chamado de **antecedente** e o conjunto de itens do lado direito da regra (Leite) é chamado de **conseqüente**.

O primeiro valor que aparece entre os parênteses corresponde à **confiança** da regra. A confiança representa a freqüência relativa (ou probabilidade condicional) entre a ocorrência do evento no conseqüente e a ocorrência do evento no antecedente. Podemos dizer que a confiança dá uma medida do poder de previsão da regra: se já soubermos que uma determinada compra inclui Pão e Manteiga, e arriscamos dizer que ela também incluirá Leite, qual será a nossa chance de acerto? Pela regra acima, a nossa chance de acerto será de 80%. Os termos confiança, freqüência relativa e probabilidade condicional podem ser usados de forma intercambiável.

O segundo valor corresponde ao **suporte** da regra. O suporte representa a freqüência de ocorrência do evento formado pela união entre o antecedente e o conseqüente da regra e dá uma medida da sua significância estatística. Na regra acima, o suporte de 50% indica que 50% de todas as transações realizadas na padaria incluíram os itens Pão, Manteiga e Leite. Regras que apresentam um suporte abaixo de um determinado valor são consideradas pouco relevantes.

Em uma tarefa de mineração de regras de associação, o usuário deve informar quais os valores mínimos de suporte e confiança que uma regra deve apresentar para ser considerada interessante. Apenas as regras que satisfazem a essas restrições são apresentadas para o usuário. (Na verdade, as regras que não satisfazem às restrições de suporte e confiança não chegam sequer a ser geradas, ao que deve ser creditada a eficiência dos algoritmos de mineração de regras de associação.)

O suporte e a confiança são o que se costuma chamar de métricas de interesse. O objetivo dessas métricas é auxiliar o usuário a identificar as regras que são mais interessantes em meio a grandes volumes de regras. Existem na literatura dezenas de

métricas de interesse (Tan et al., 2002). Algumas são mais intuitivas, como o suporte e a confiança; outras são resultado de cálculos matemáticos complexos, praticamente impossíveis de serem expressas em palavras. Existem métricas que são mais adequadas para domínios de aplicação específicos, outras para bases de dados de natureza específica. A escolha da melhor métrica de interesse a ser usada e das faixas de valores interessantes para uma dada métrica de interesse não é uma escolha trivial.

Uma tarefa de mineração de regras de associação pode ser resumida através dos seguintes passos. O usuário seleciona os dados que deseja minerar e fornece valores mínimos para o suporte e para a confiança. O algoritmo gera as regras que apresentarem suporte e confiança acima dos valores mínimos especificados. Mas a quantidade de regras geradas é tipicamente muito grande, podendo chegar à casa das centenas de milhares. O usuário então se utiliza das outras métricas de interesse para encontrar as regras mais interessantes no conjunto das regras geradas. Cada regra selecionada é analisada pelo usuário, que deve transformar a informação que lhe é apresentada em conhecimento, diretamente aplicável no seu dia-a-dia.

A mineração de regras de associação pode ser usada para suportar diferentes tipos de tarefas. A mais comum delas é a análise exploratória de dados. Neste tipo de tarefa, os usuários não sabem o que estão procurando quando começam a utilizar o sistema. A idéia é que o sistema revele ao usuário padrões que possam ser interessantes para ele, e que a partir desses padrões o usuário possa formular hipóteses, que vão sendo refinadas e levam a novas hipóteses, até que ele consiga obter uma informação interessante. A mineração de regras de associação pode ser usada também para o teste de hipóteses. Neste caso, o usuário parte de uma hipótese e tenta encontrar as regras que confirmem ou refutem tal hipótese.

Uma outra tarefa que a mineração de regras de associação pode suportar é a previsão. Vimos que a confiança de uma regra dá uma medida do seu poder de previsão. Assim, se observarmos que o evento representado no antecedente de uma regra ocorreu, podemos prever o evento representado no conseqüente da mesma com uma probabilidade de acerto que equivale à confiança da regra.

Finalmente, podemos utilizar a mineração de regras de associação em uma tarefa conhecida como classificação parcial. Neste caso, estamos interessados apenas nas regras que tiverem um único item no conseqüente, sempre associado a um determinado atributo (classe-alvo), e utilizamos essas regras para prever o valor nesse atributo com base nas informações conhecidas, ou seja, aquelas representadas pelo antecedente da regra. A classificação parcial é uma caso particular da tarefa de previsão.

1.2 Motivação

A mineração de regras de associação possui diversas aplicações, que extrapolam em grande medida a sua aplicação original. Se, no seminal artigo de 1993, Agrawal et al. propuseram esta técnica para que ela fosse utilizada na análise dos hábitos de compra dos clientes de um supermercado, hoje esta mesma técnica é utilizada para as mais diferentes aplicações, incluindo:

- detecção de fraudes em seguros e compras governamentais;
- análise exploratória em ciências sociais;
- *database marketing* (generalização da análise do carrinho de compras);
- análise de risco em operações de crédito;
- detecção de intrusos em redes de computadores;
- pesquisas médicas e farmacêuticas e
- recuperação da informação.

Embora os sistemas de mineração de regras de associação venham se popularizando desde o seu aparecimento no início da década de 90, eles normalmente ainda requerem a presença de um especialista, independente do domínio em que sejam aplicados (Soukup e Davidson, 2002). Esta dependência faz com que esses sistemas sejam muitas vezes sub-utilizados, já que são poucos os usuários que podem contar com a ajuda de um consultor especializado.

Desenvolver interfaces de boa qualidade para sistemas de mineração de dados é um novo desafio tanto para a área de interação humano-computador (IHC), quanto para a área de mineração de dados (Fayyad et al., 2001; Shneiderman, 2002). Em particular, o desenvolvimento de interfaces para sistemas de mineração de regras de associação apresenta duas grandes dificuldades para o projetista: primeiro, a quantidade de regras geradas é tipicamente muito grande, o que dificulta a sua exibição e torna necessária a criação de mecanismos que permitam ao usuário encontrar as regras mais interessantes; segundo, os conceitos relacionados à técnica são complexos, o que torna difícil explicá-los ao usuário.

O surgimento das técnicas de mineração de dados foi motivado principalmente pela necessidade de se analisarem grandes volumes de dados e pela incapacidade das técnicas tradicionais em lidar com esses grandes volumes. No entanto, a quantidade de dados a ser minerada tem se tornado tão grande que o volume de padrões detectados pelos algoritmos de mineração de dados é ainda muito grande, criando um problema de

mineração de segundo nível. Uma única tarefa de mineração de regras de associação, por exemplo, pode gerar centenas de milhares de regras. Os projetistas enfrentam então o seguinte problema: como exibir essa grande quantidade de regras para o usuário e como ajudá-lo a encontrar as regras que são mais interessantes? Existem diversas propostas para lidar com esse problema: algumas são algorítmicas, outras utilizam técnicas de IHC. A nossa contribuição para a solução desse problema se encontra nesta segunda categoria.

O segundo problema enfrentado pelos projetistas dessas interfaces é que os conceitos relacionados à mineração de regras de associação são complexos, e geralmente não fazem parte do domínio do usuário. Esses conceitos envolvem o próprio significado de regra de associação, as métricas de interesse, os relacionamentos entre as regras e o processo de geração de regras. Cabe aos projetistas transmitir ao usuário os conhecimentos específicos de mineração de regras de associação que são necessários para que ele possa utilizar o sistema. As soluções atuais não abordam esse problema, com a notável exceção do trabalho de Hofmann et al. (2000).

1.3 Objetivos e Contribuições

O objetivo por trás desta dissertação é o desenvolvimento de uma interface que permita a interação do usuário com o conjunto de regras geradas em uma tarefa de mineração de regras de associação. Por interação, queremos dizer a visualização das regras, a seleção das regras mais interessantes e o entendimento tanto do significado de uma regra quanto de todos os conceitos relacionados a regras de associação, sem os quais o usuário não é capaz de utilizar o sistema satisfatoriamente. Assim, estamos interessados em abordar os dois problemas citados na seção anterior.

As principais contribuições do nosso trabalho são:

- Desenvolvimento e avaliação de duas estratégias para a visualização de regras de associação;
- Identificação dos aspectos que devem ser compreendidos pelo usuário para que ele possa utilizar satisfatoriamente um sistema de mineração de regras de associação e os custos atrelados à não compreensão de cada um desses aspectos, e
- Caracterização e análise do comportamento dos usuários neste domínio.

1.4 Metodologia

O foco deste trabalho concentra-se principalmente no usuário. O nosso objetivo é o desenvolvimento de uma interface de mineração de regras de associação que seja a mais adequada possível à utilização pelos usuários, sejam eles especialistas em mineração de dados ou não. Portanto, não poderíamos adotar como metodologia de desenvolvimento uma abordagem que não fosse centrada no usuário. Gould e Lewis (1985) recomendam três princípios que devem guiar o *design* de qualquer sistema que esteja sendo desenvolvido para ser usado por pessoas:

- Primeiro, os projetistas devem entender quem serão seus usuários. Este entendimento só pode ser obtido através do estudo das características dos usuários e da natureza do trabalho a ser realizado por eles;
- Segundo, os usuários devem ser envolvidos o mais cedo possível no processo de desenvolvimento. Eles devem simular a realização do seu trabalho real através de protótipos, e o seu desempenho e reações devem ser observadas, registradas e analisadas, e
- Terceiro, quando problemas são encontrados nos testes com usuários, eles devem ser corrigidos. Isto significa que o projeto deve ser interativo: deve haver um ciclo de desenho, teste, avaliação e redesenho, repetido quantas vezes for necessário.

Esses princípios representam a essência do conceito de Desenho Centrado no Usuário e foram aplicados no desenvolvimento das nossas interfaces. Começamos com uma análise de usuários e tarefas e com o desenvolvimento de uma primeira versão. Esta versão foi avaliada pelos usuários, e desta avaliação foram tiradas as conclusões que guiaram o desenho da segunda versão, a qual foi novamente avaliada pelos usuários. Este ciclo se repetiu até que chegássemos à versão atual.

Como já dissemos, um dos principais problemas de interação humana em sistemas de mineração de regras de associação está relacionado ao entendimento por parte dos usuários dos conceitos específicos da técnica, sem os quais ele não consegue utilizar o sistema de maneira satisfatória. A teoria da engenharia semiótica (de Souza, 2005) entende um sistema computacional como sendo um artefato intelectual gerado pelo projetista do sistema. Este artefato resulta do seu entendimento sobre o perfil, contexto e necessidades dos usuários, e de suas decisões sobre que problemas eles querem resolver e como eles podem fazê-lo. A interface do sistema é responsável por transmitir aos usuários estas decisões. Assim, ela é vista como uma mensagem do projetista para os usuários sobre suas decisões e princípios de interação que guiaram seu *design*. Esta mensagem caracteriza uma meta-comunicação, uma vez que os usuários devem

entendê-la à medida que trocam mensagens com o sistema. Mais do que isto a interface é entendida como o preposto do projetista, uma vez que ela "fala" em seu nome com o usuário. Para caracterizar a qualidade desta comunicação entre o projetista e usuário, a engenharia semiótica propõe a propriedade de comunicabilidade (Prates et al., 2000; de Souza, 2005). Quando o projetista consegue realizar esta comunicação através da interface de forma eficiente, pode-se dizer que a interface tem alta comunicabilidade. A engenharia semiótica argumenta que para um usuário utilizar um sistema computacional com eficiência ele deve conseguir entender a visão do projetista que guiou o seu projeto e desenvolvimento, em outras palavras o sistema deve ter alta comunicabilidade.

Utilizamos os princípios da engenharia semiótica para estudar os problemas de entendimento enfrentados pelo usuário em ambientes de mineração de dados. Nesses ambientes, para se ter uma alta comunicabilidade é preciso que o usuário entenda não apenas a decisão do projetista de utilizar a técnica de mineração de dados, mas também conceitos e aspectos da própria técnica. Desta forma, estes ambientes apresentam um desafio a mais de comunicabilidade para seus projetistas, uma vez que eles devem comunicar não apenas soluções relativas ao que se pode fazer no domínio do usuário e como fazê-lo, mas conhecimentos técnicos específicos à área de mineração de dados, que normalmente os usuários não possuem. Com o objetivo de apoiar os projetistas destes ambientes ao lidarem com este desafio, neste trabalho caracterizamos que aspectos sobre este conhecimento técnico são importantes de se comunicar ao usuário, os desafios de fazê-lo e os impactos para o usuário quando isto não acontece. Para isto nos baseamos na literatura de interação em sistemas de mineração de regras de associação e em dados obtidos a partir de nossas observações no uso do sistema.

1.5 Organização da Dissertação

O restante desta dissertação está organizado da seguinte forma. No próximo capítulo, fazemos uma revisão da literatura relacionada ao nosso trabalho. No Capítulo 3, explicamos as estratégias que desenvolvemos para a visualização de regras associação. Em seguida, no Capítulo 4, apresentamos as resultados que obtivemos nas nossas avaliações com usuários. Finalmente, no Capítulo 5, apresentamos as nossas conclusões e apontamos possíveis direções para trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

Neste capítulo, apresentamos inicialmente um resumo da área de visualização da informação. Em seguida, mostramos como a área de visualização da informação se combina com a mineração de dados, na chamada mineração visual de dados. Por fim, apresentamos os trabalhos relacionados a visualização de regras de associação.

2.1 Visualização da Informação

A utilização de elementos gráficos para representar e transmitir idéias não é uma prática recente. Desde os longínquos mapas cartográficos, passando pelos gráficos de William Playfair, economista escocês do século dezoito, até chegar ao trabalho do renomado estatístico de Princeton, John W. Tukey, nos anos de 1960, que a comunicação de idéias complexas com clareza, precisão e eficiência vem provocando a imaginação e a criatividade daqueles que se dispõem a tal tarefa.

Em 1869 o engenheiro francês Charles Joseph Minard concebeu o que mais tarde seria considerado por Edward Tufte em seu aclamado *The Visual Display of Quantitative Information* "o melhor gráfico estatístico já desenhado" (Figura 2.1). O gráfico, uma combinação de mapa e série temporal, narra de forma magistral a campanha do exército de Napoleão Bonaparte na Rússia em 1812. Começando na esquerda, a faixa mais clara mostra a evolução das tropas em solo russo da fronteira polonesa até Moscou, no extremo direito, com a largura da faixa representando o tamanho do exército em cada ponto do mapa. A faixa escura mostra a retirada das tropas de Moscou, com destaque para as sucessivas reduções na largura da faixa, que associadas à escala de temperatura na parte inferior do mapa ilustram as baixas impostas ao contingente de homens pelo rigor do inverno russo. A forma extremamente elegante com que tantas informações foram representadas em um único gráfico serve de inspiração até os nossos dias.

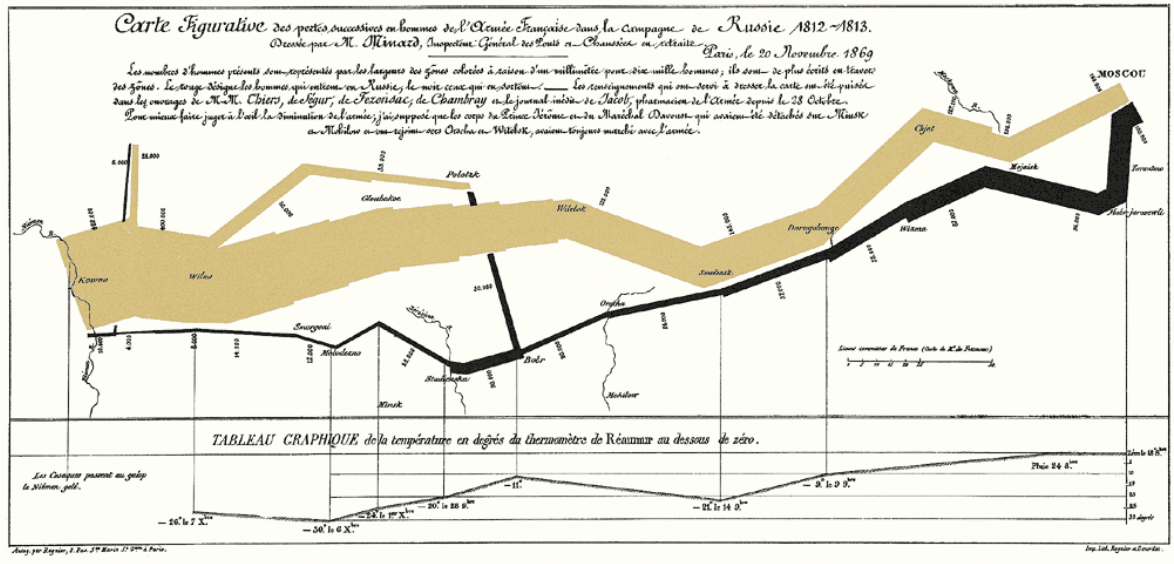


Figura 2.1: Gráfico de Charles Joseph Minard ilustrando a campanha do exército de Napoleão na Rússia em 1812 (Tufte, 2005).

Especialmente depois de Playfair, a quem é creditada a criação dos princípios fundamentais dos gráficos modernos, os gráficos estatísticos passaram a exercer um papel central na política e na economia, nas corporações e no meio acadêmico, facilitando a compreensão dos dados e auxiliando no processo de tomada de decisão. Mas o melhor ainda estava por vir: o surgimento e a evolução explosiva dos computadores colocaram à disposição dos analistas de dados um exuberante conjunto de novas e poderosas ferramentas, como planilhas eletrônicas, pacotes estatísticos e sistemas de suporte à decisão. E os gráficos – agora muito mais dinâmicos, interativos e ricos em cores – se consolidaram como os componentes essenciais de todas essas ferramentas.

Visualização da informação é uma nova área de pesquisa, que foca no uso de técnicas gráficas para ajudar as pessoas a entender e analisar diversos tipos de dados. Enquanto a visualização científica envolve a apresentação de dados que têm alguma correspondência física ou geométrica, a visualização da informação foca em dados abstratos que não possuem tal correspondência, como dados simbólicos, tabulares, hierárquicos, textuais ou redes. As pesquisas em visualização da informação se dividem em duas linhas básicas. A primeira foca em aspectos relacionados à apresentação de informações estáticas, se preocupando basicamente com questões como expressividade, percepção e estética. Como principais expoentes desta primeira linha, podemos citar Bertin (1983), Tufte (1986, 1997, 1990) e Cleveland (1985, 1993). Este último, por exemplo, comparou empiricamente a precisão de diferentes propriedades gráficas para representar informações quantitativas (Figura 2.2), demonstrando que a posição é a propriedade mais precisa. O trabalho de Cleveland foi mais tarde estendido por Mackinlay (1986) para contemplar

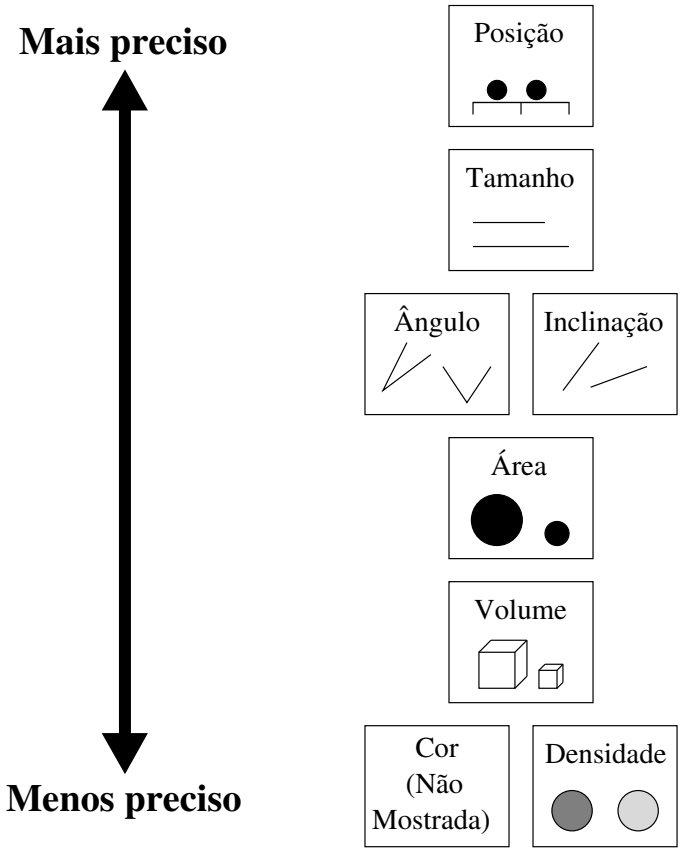


Figura 2.2: Ranking de precisão das propriedades gráficas.

também informações qualitativas, não tendo sido no entanto avaliado empiricamente.

A segunda linha foca em aspectos interativos, como *zoom*, filtros, navegação, manipulação direta, ajuste dinâmico das informações ao espaço disponível, distorção, oclusão etc. Aqui destaca-se o trabalho de Shneiderman, North, Plaisant, Inselberg e tantos outros, responsáveis pela invenção de admiráveis ferramentas de exploração da informação, como *dynamic queries*, *starfield displays*, o projeto *Visible Human Explorer*, *treemaps* e coordenadas paralelas.

Embora muita pesquisa venha sendo realizada nesta área, a criação de visualizações tem sido tratada ainda como um processo essencialmente *ad-hoc*, sem qualquer método formal. Existem algumas tentativas de formalização do processo, mas nenhuma conseguiu ganhar uma aceitação muito ampla até o momento. A utilização de tais modelos formais traria três grandes benefícios para a área. Em primeiro lugar, eles poderiam oferecer ao projetista um direcionamento consistente sobre como abordar o processo de criação de visualizações. Em segundo lugar, eles poderiam ajudar de alguma forma na automatização total ou parcial do processo de criação de visualizações. E finalmente eles poderiam prover uma base objetiva para a comparação da efetividade de diferentes visualizações na realização de uma determinada tarefa, oferecendo também inspiração

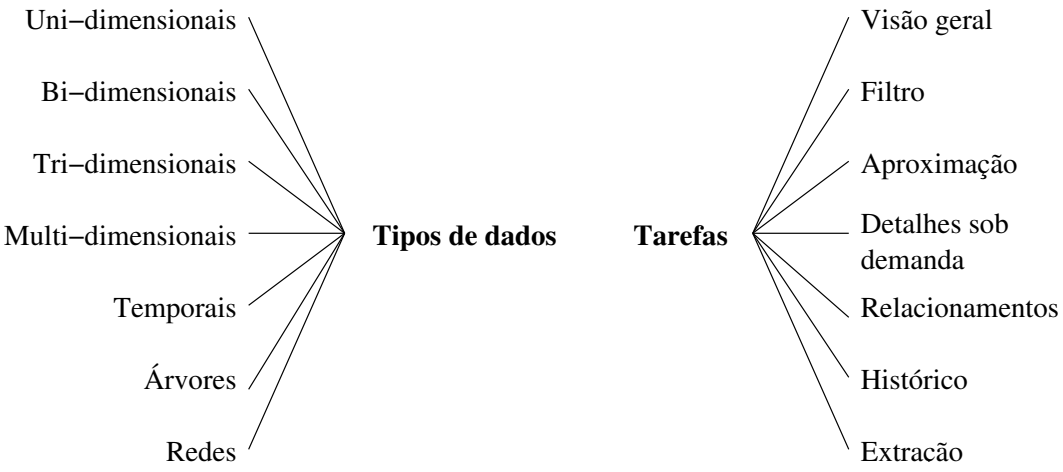


Figura 2.3: Taxonomia para visualização da informação baseada em tipos de dados e tarefas.

para a criação de novas técnicas.

Existem, no entanto, diversas propostas de taxonomias para visualização da informação. Shneiderman, por exemplo, propõe uma taxonomia baseada em tipos de dados e tarefas (Figura 2.3) (Shneiderman, 1996). Os **tipos de dados** são sete: uni-, bi- e tri-dimensionais, multi-dimensionais, temporais, redes e árvores. Esses tipos de dados caracterizam os objetos de informação e são organizados de acordo com os problemas que os usuários estão tentando resolver. Por exemplo: em informações bi-dimensionais, tais como mapas, os usuários estão tentando entender as adjacências ou navegar nos caminhos possíveis; em informações estruturadas em árvores, os usuários estão tentando entender os relacionamentos pai-filho-irmão. As **tarefas** são as ações que os usuários querem realizar, em um nível mais alto de abstração: visão-geral (obter uma visão geral da coleção inteira), aproximação (focalizar nos itens de interesse), filtro (filtrar os itens que interessam), detalhes-sob-demanda (selecionar um item e obter detalhes quando necessário), relacionamentos (visualizar os relacionamentos entre os itens), histórico (manter um histórico de ações para suportar operações de "desfazer", "refazer" e refinamento progressivo) e extração (permitir extração de subconjuntos de itens e dos parâmetros de consulta).

Existem ainda outras taxonomias, como a taxonomia proposta por Card e Mackinlay (1997), que divide o campo de visualização da informação em várias categorias com base no tipo de dados; ou a taxonomia proposta por Chi (2000), que se baseia não apenas nos tipos de dados, mas também nos operadores de processamento inerentes a cada técnica de visualização. Essas taxonomias são úteis porque ajudam os projetistas a identificar rapidamente várias técnicas que podem ser aplicadas ao seu domínio de interesse, e também entender como aplicar e implementar essas técnicas.

Acreditamos que os estudos no campo de visualização da informação podem fornecer alguns dos fundamentos teóricos necessários para o desenvolvimento de interfaces de mineração de dados mais eficientes. A maior parte das decisões de projeto que fizemos neste trabalho se apoiaram em resultados obtidos nessa área.

2.2 Mineração Visual de Dados

Uma tendência recente é a combinação de técnicas de mineração de dados com técnicas de visualização da informação. Embora as áreas de mineração de dados e visualização da informação remontem a um passado comum e compartilhem alguns dos seus objetivos básicos, elas seguiram linhas de pesquisa completamente distintas, e só agora parecem caminhar rumo a uma convergência. Esta tendência pode ser comprovada por algumas publicações recentes, que reivindicam a interação entre as áreas, sob o argumento de que elas podem se beneficiar mutuamente nesta interação.

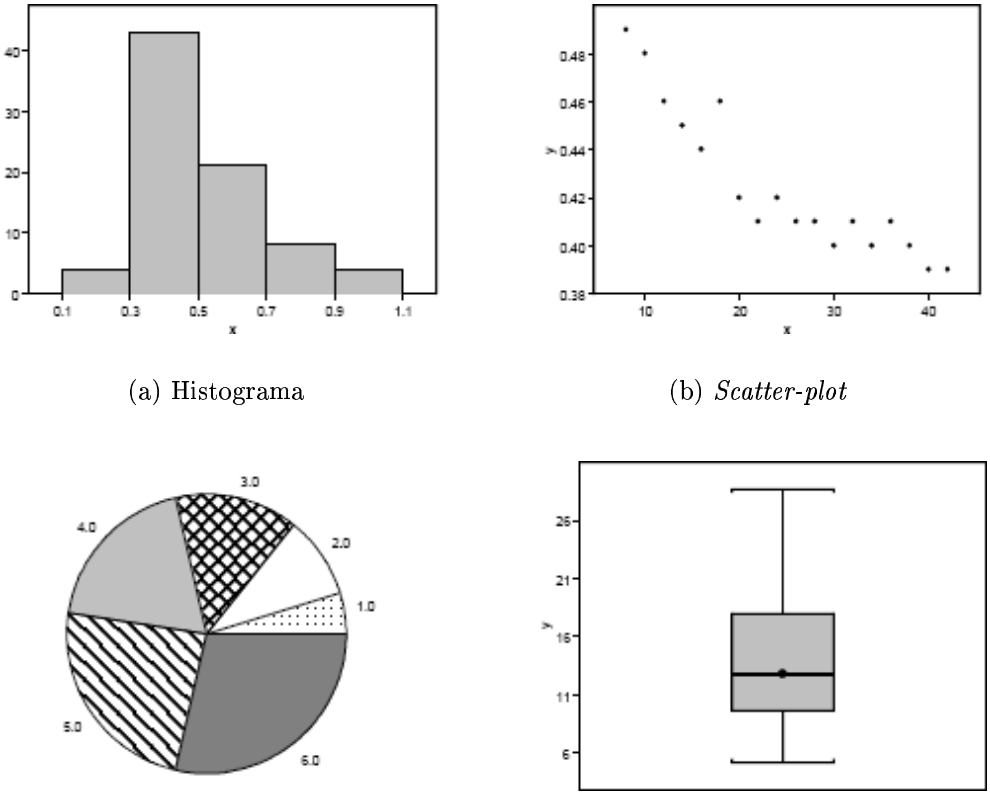
O objetivo da visualização da informação, assim como da mineração de dados, é revelar padrões e anomalias presentes nos dados. Enquanto na mineração de dados isso é feito com o uso de algoritmos – como algoritmos de agrupamento ou de mineração de regras de associação – na visualização da informação a idéia é que o usuário possa encontrar os padrões e anomalias apenas visualizando as representações gráficas dos dados. A esta tarefa de "mineração" realizada visualmente pelos usuários costuma-se chamar de mineração visual de dados.

Alguns autores, no entanto, utilizam o termo mineração visual de dados sempre que técnicas de visualização são combinadas com a mineração de dados. Isto pode ocorrer de quatro maneiras distintas:

- Visualização de dados;
- Visualização do processo de descoberta de conhecimento;
- Mineração interativa, e
- Visualização de modelos.

A primeira consiste na visualização dos dados em seu formato original, com o objetivo de identificar algum tipo de padrão ou anomalia. A visualização de dados se confunde com o conceito de mineração visual de dados discutido anteriormente e com o próprio conceito de visualização da informação. Outras tarefas associadas à visualização de dados incluem a análise da distribuição dos dados e a análise de parâmetros estatísticos, como média, mediana e variância. Dentre as ferramentas para análise de dados, destacam-se os gráficos estatísticos tradicionais, como histogramas, *scatter-plots*,

gráficos de pizza, *box-plots*, gráficos de linha etc. (Figura 2.4), além de ferramentas típicas de visualização da informação, como *starfield displays*, *treemaps*, coordenadas paralelas etc.



(a) Histograma

(b) *Scatter-plot*

(c) Gráfico de pizza

(d) *Box-plot*

Figura 2.4: Exemplos de visualização de dados.

A segunda maneira pela qual a visualização da informação é empregada na mineração de dados é a visualização do processo de descoberta de conhecimento em bancos de dados. Vimos anteriormente que o processo de descoberta de conhecimento em bancos de dados é composto de quatro etapas. A visualização pode ajudar o usuário a identificar de onde vieram os dados, quais os tipos de preparação que eles sofreram, que algoritmo de mineração foi utilizado e onde os resultados foram armazenados (Figura 2.5).

A terceira maneira se refere à interação do usuário durante a etapa de mineração, quando ele visualiza os resultados parciais do algoritmo, podendo interferir na execução do mesmo, seja alterando os parâmetros, seja fornecendo *feedback* para direcionar a atuação do algoritmo. Aggarwal (1998), por exemplo, propõe uma cooperação entre o usuário e o computador no processo de construção de uma árvore de decisão, em tarefas de classificação. Na sua proposta, o usuário se vale de informações sobre a distribuição

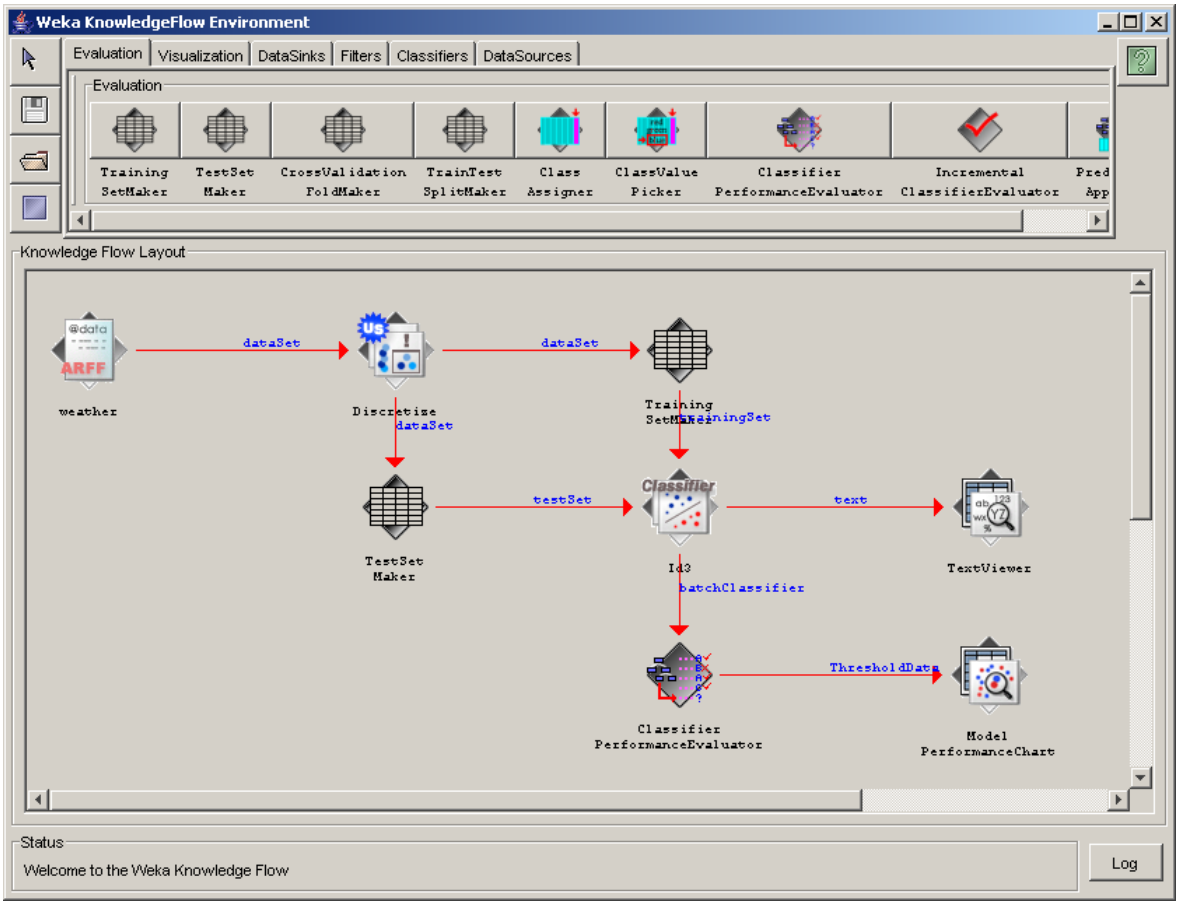


Figura 2.5: Exemplo de visualização do processo de descoberta de conhecimento em banco de dados.

dos dados em um determinado atributo para decidir o melhor ponto de divisão para esse atributo durante a construção da árvore. Ankerst et al. (2000) propõem uma técnica para análise de agrupamentos na qual o usuário contribui com o algoritmo informando se os grupos que estão sendo gerados são bem distintos (situação desejada) ou não. Essa decisão é difícil para o algoritmo, mas para o usuário ela se torna fácil a partir da visualização dos grupos gerados.

A quarta e última forma de combinação entre mineração de dados e visualização da informação ocorre na visualização dos modelos gerados em uma tarefa de mineração de dados. Aqui, não são mais os dados originais que são visualizados, mas os padrões e modelos derivados desses dados em uma tarefa de mineração de dados, como agrupamentos hierárquicos e árvores de decisão. O nosso trabalho se encontra nesta categoria, pois estamos interessados na visualização de regras de associação, que são o resultado de uma tarefa de mineração de dados.

2.3 Visualização de Regras de Associação

Conforme observado por Hofmann et al. (2000), um sistema de visualização de regras de associação deveria (1) ajudar os usuários a encontrar as regras mais interessantes e (2) ajudá-los a entender essas regras. O primeiro problema é causado principalmente pela enorme quantidade de regras geradas em uma tarefa típica de mineração de regras de associação, e já foi abordado tanto com um enfoque algorítmico quanto com um enfoque em IHC. O outro problema é causado pela complexidade inerente às regras de associação, sendo essencialmente um problema de IHC. Nesta seção, vamos rever os trabalhos que tratam de cada um desses problemas.

2.3.1 Visualizando as Regras

O problema de identificar as regras mais interessantes no conjunto de regras que podem ser geradas em uma tarefa de mineração de regras de associação já foi abordado de várias formas. Algumas técnicas permitem que o usuário defina padrões para as regras que ele julga interessantes, utilizando expressões regulares; o sistema retorna então apenas as regras que casam com os padrões definidos pelo usuário (Klemettinen et al., 1994). Outras, ao contrário, permitem que o usuário informe as correlações que ele já sabe que existem na base; o sistema retorna então apenas aquelas regras que são inesperadas se comparadas com esse conhecimento prévio do usuário (Silberschatz e Tuzhilin, 1996). Existem ainda as técnicas que tentam eliminar as redundâncias existentes entre as regras ou sumarizar o conjunto de regras geradas, exibindo apenas aquelas que não podem ser derivadas de outras (Liu et al., 1999).

Alguns trabalhos investigam a utilização de outros tipos de restrições para as regras a serem geradas, que vão além dos já tradicionais suporte e confiança mínimos: é a chamada mineração baseada em restrições (Aggarwal e Yu, 1998; Bayardo et al., 2000). A idéia é utilizar todas as restrições possíveis de forma a reduzir ao máximo o número de regras geradas, diminuindo também o tempo de execução do algoritmo. No entanto, pela natureza exploratória da mineração de regras de associação, a tendência é que sejam necessárias várias minerações sucessivas até que se chegue ao resultado desejado. De acordo com Goethals e den Bussche (1999), esta situação acaba sendo pior do ponto de vista do custo de execução do que realizar uma mineração menos restritiva e que gere mais resultados de uma só vez.

Todos esses trabalhos têm um enfoque essencialmente algorítmico, embora possam ser suportados por interfaces gráficas. Estamos mais interessados nos trabalhos que propõem a utilização de técnicas de visualização para ajudar na identificação das regras mais interessantes. Podemos agrupar esses trabalhos basicamente em duas ca-

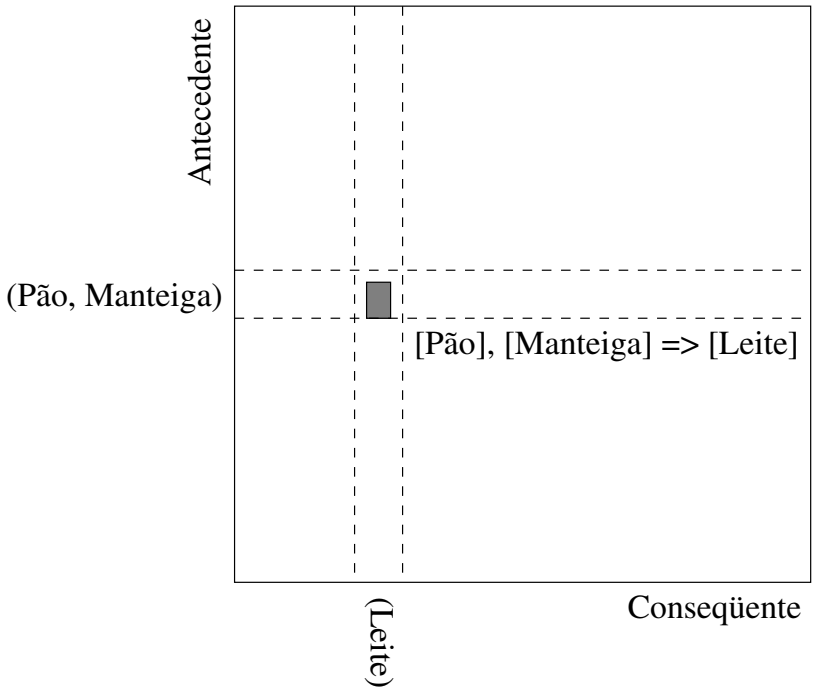
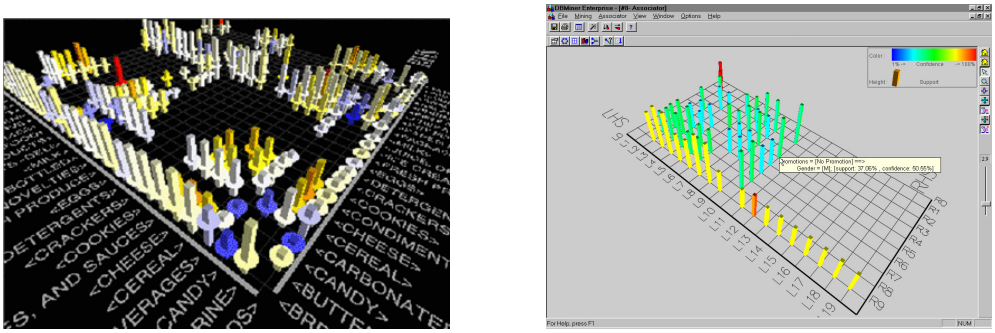


Figura 2.6: Matriz antecedente×conseqüente.

tegorias, de acordo com o paradigma visual empregado na exibição do conjunto de regras. Na primeira categoria encontram-se as técnicas que organizam o conjunto de regras na forma de uma matriz antecedente×conseqüente; e na segunda as técnicas que o organizam na forma de um grafo.

A maneira mais popular de organizar as regras em matriz é distribuir os conjuntos de itens ao longo de dois eixos perpendiculares, um deles representando o antecedente e o outro representando o conseqüente (Figura 2.6). As regras são então representadas nas células localizadas na interseção entre os respectivos antecedentes e conseqüentes. Para representar as regras são usadas formas geométricas, cujas propriedades gráficas (cor, tamanho, textura etc.) são utilizadas para representar as diferentes métricas de interesse associadas à regra. A maioria dos pacotes comerciais de mineração de dados oferecem implementações para esta técnica, como o Mineset, o IBM Intelligent Miner, o SAS Enterprise Miner e o DBMiner. A Figura 2.7 traz dois exemplos de sistemas que utilizam esta abordagem. Em ambos, as regras são representadas por barras verticais, com a cor da barra representando a confiança e a altura representando o suporte (na Figura 2.7a, é usado ainda um disco para representar uma terceira métrica, cujo valor é dado pela posição vertical do disco na barra). O principal problema com esta abordagem é que a visualização não escala bem tanto para grandes volumes de regras, quanto para regras com muitos itens no antecedente ou no conseqüente.

Wong et al. (1999) tentam resolver o problema da grande quantidade de itens no



(a) IBM Intelligent Miner (b) DBMiner

Figura 2.7: Exemplos de ferramentas que utilizam a abordagem baseada em uma matriz antecedente×conseqüente.

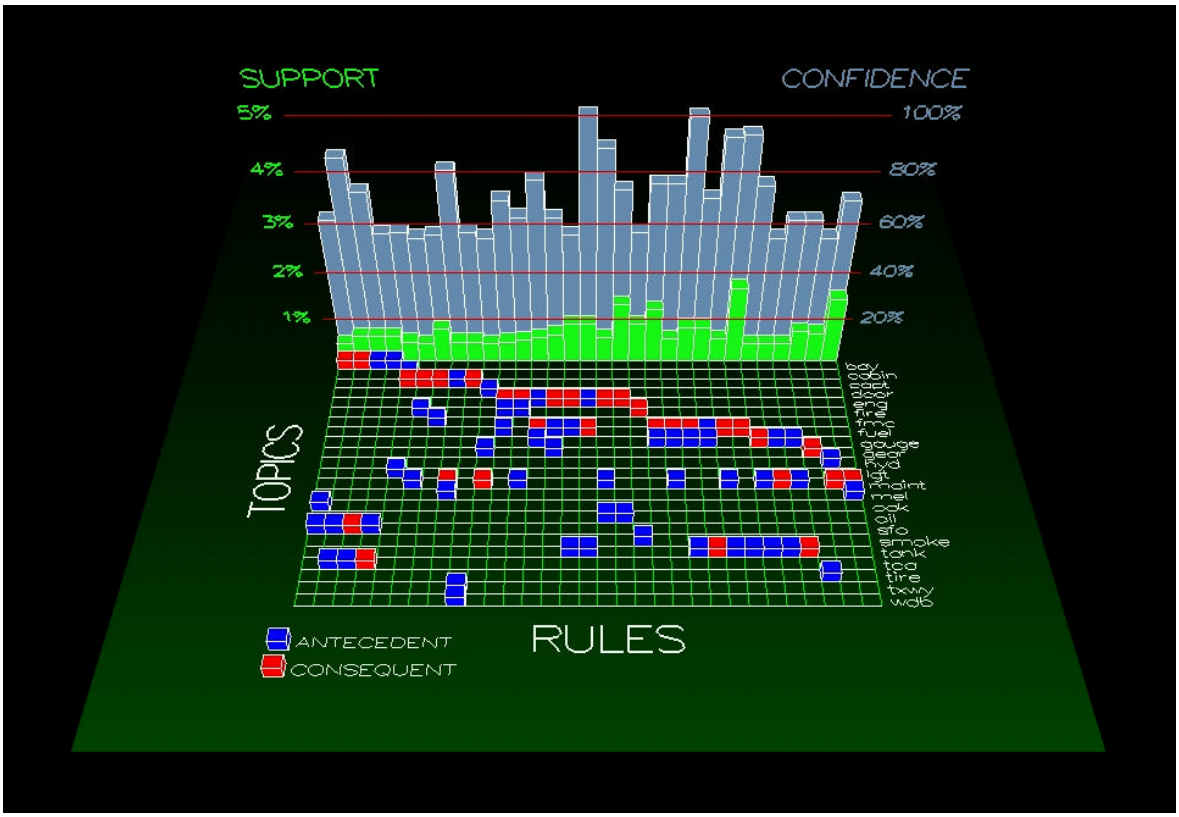


Figura 2.8: Abordagem proposta por Wong et al.

antecedente ou no conseqüente com uma abordagem um pouco diferente (Figura 2.8). Na sua abordagem, as regras são representadas em uma matriz, onde as colunas representam as regras e as linhas representam os itens das regras. Para diferenciar os itens que estão no antecedente dos itens que estão no conseqüente de uma regra, são usadas duas cores: azul para representar os itens do antecedente e vermelho para representar os itens do conseqüente. Assim, podem ser representadas facilmente regras com tantos

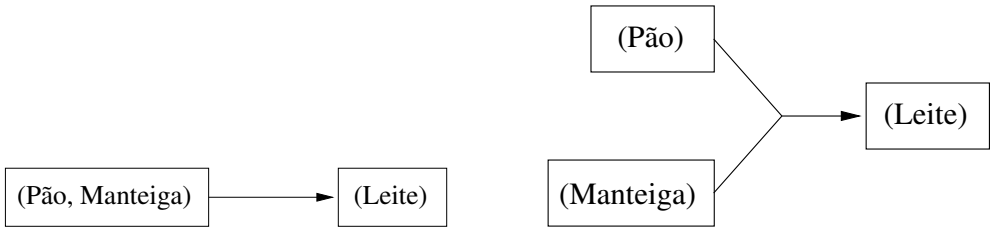
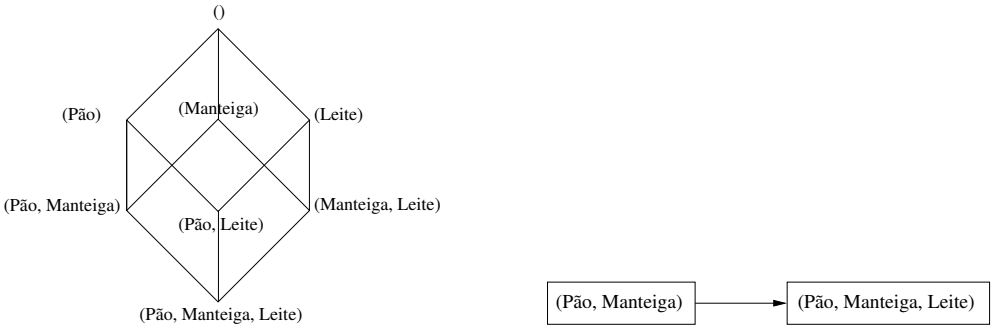


Figura 2.9: Abordagem baseada em grafos, em duas variações possíveis.



(a) *Itemset lattice*

(b) Grafo baseado no *itemset lattice*

Figura 2.10: Uma outra variação da abordagem baseada em grafos.

itens quantos forem as linhas na matriz. No entanto, embora eles tenham resolvido em parte o problema, eles criaram uma dificuldade adicional de leitura das regras, já que o usuário tem que ficar consultando a legenda várias vezes para identificar quais itens estão no antecedente e quais estão no conseqüente.

As técnicas que utilizam grafos para organizar o conjunto de regras são também bastante populares (Zaki e Phoophakdee, 2003; Hao et al., 2001; Kuntz et al., 2000; Rainsford e Roddick, 2000). Nessas técnicas, os nodos do grafo representam os itens ou conjuntos de itens e as arestas representam as regras, com o nodo origem representando o antecedente e o nodo destino representando o conseqüente da regra (Figura 2.9). Por exemplo, a regra AB $\Rightarrow$ C é representada pela aresta que une o nodo AB ao nodo C (em uma pequena variação, esta regra seria representada pela união entre a aresta que une o nodo A ao nodo C e a aresta que une o nodo B ao nodo C). As métricas de interesse são representadas por atributos das arestas, tais como cor ou espessura.

Alguns trabalhos organizam o grafo como um subconjunto do *itemset lattice* (Figura 2.10a). O *itemset lattice* representa todas as combinações (ou conjuntos) possíveis de itens na base de dados: no primeiro nível é representado o conjunto vazio, no segundo nível são representados todos os conjuntos de itens de tamanho 1, no terceiro

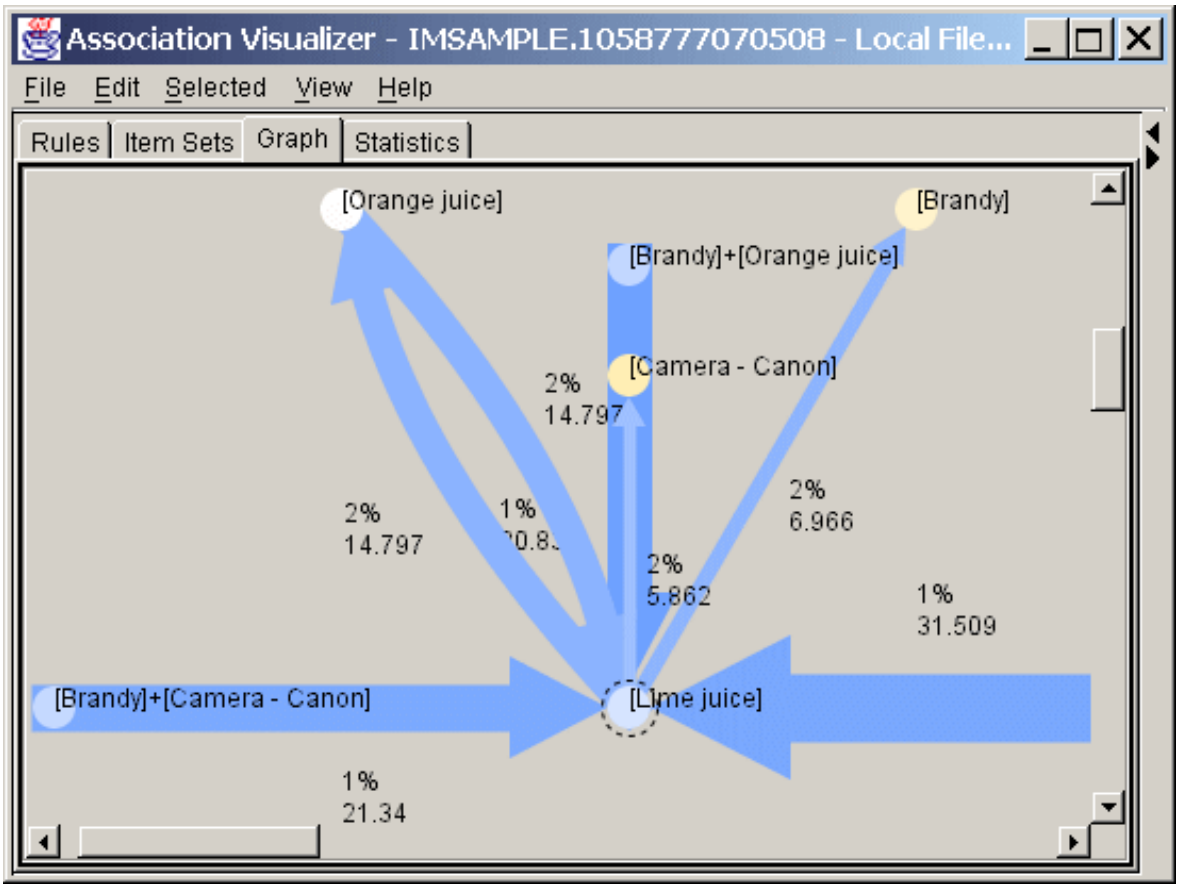


Figura 2.11: Exemplo de ferramenta que utiliza uma abordagem baseada em grafos. A largura da aresta representa a confiança da regra.

nível todos os conjuntos de itens de tamanho 2 e assim por diante. Uma transição de um nodo de um determinado nível para um nodo do nível imediatamente inferior representa a adição de um item ao conjunto representado no nodo origem, que resulta no conjunto representado no nodo destino. As abordagens baseadas em grafos que utilizam o conceito de *itemset lattice* representam as regras através dessas transições, onde o antecedente da regra é dado pelo conjunto representado no nodo origem e o conseqüente é dado pelo item que é adicionado na transição para o nodo destino. Neste caso, a regra AB $\Rightarrow$ C seria representada pela aresta que une o nodo AB ao nodo ABC (Figura 2.10b) (Kuntz et al., 2000). O *itemset lattice* pode ser visto como uma forma de organizar as regras hierarquicamente, onde descer no *itemset lattice* representa uma operação de especialização e subir no *itemset lattice* representa uma operação de generalização.

As abordagens baseadas em grafos funcionam bem para uma quantidade pequena de regras (Figura 2.11). Mas para uma quantidade um pouco maior elas não escalam, provocando um efeito indesejado (Figura 2.12). Além disso, nas abordagens baseadas em grafos a identificação das regras mais interessantes é dificultada, principalmente

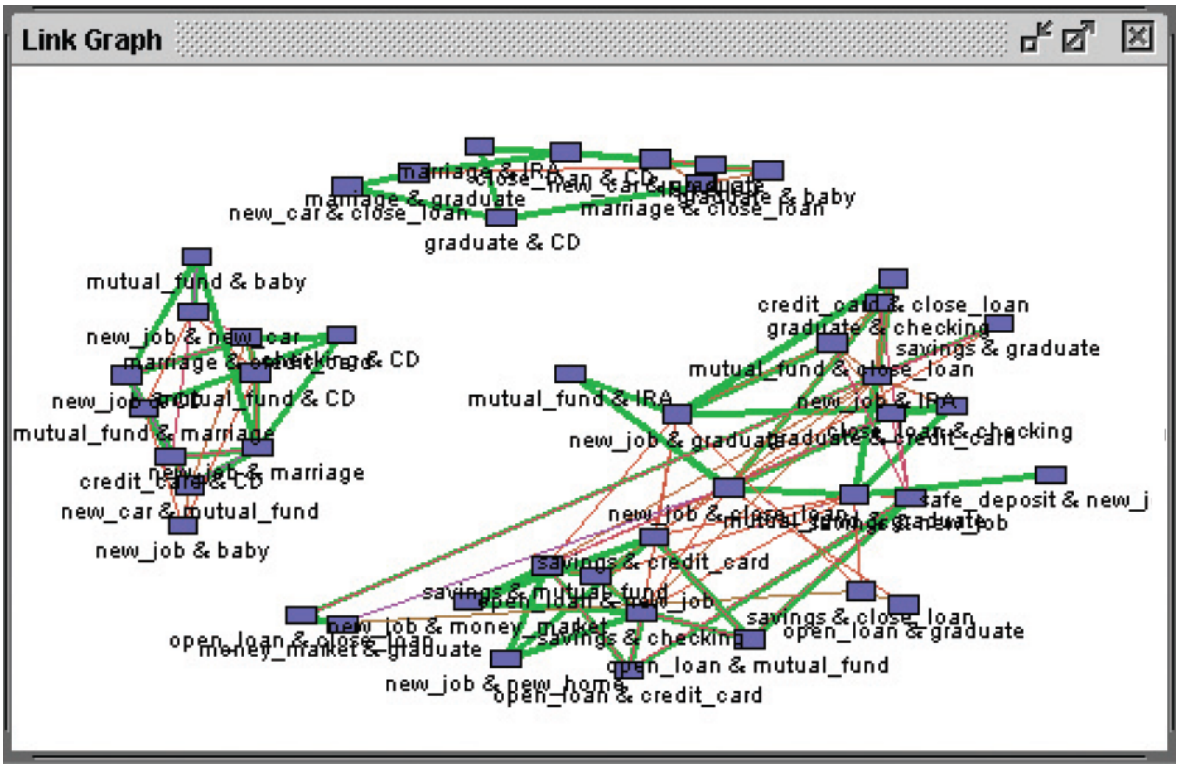
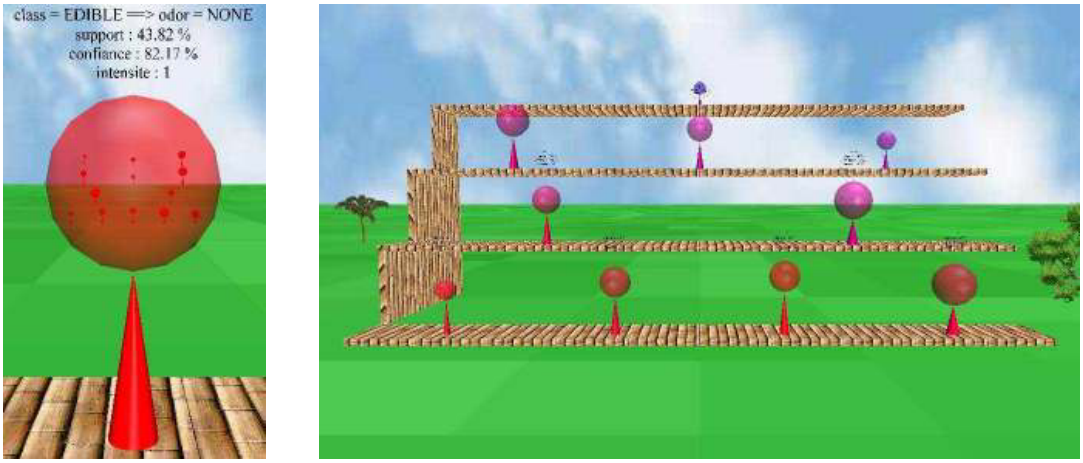


Figura 2.12: Uma outra ferramenta baseada em grafos. Quando o número de regras é muito grande, fica difícil distinguir as arestas.

porque a representação gráfica das métricas de interesse nas arestas não é muito flexível, se limitando às propriedades de largura e cor da aresta.

Blanchard et al. (2003) propõem uma metáfora visual diferente para representar regras de associação (Figura 2.13). Na sua proposta, as regras são representadas por esferas colocadas sobre cones, com a área da esfera representando o suporte e a altura do cone representando a confiança. As regras são espalhadas em duas arenas, com a posição da regra sendo usada para representar uma terceira métrica de interesse: as regras mais surpreendentes de acordo com esta métrica são posicionadas mais embaixo e à frente na arena; as regras menos surpreendentes são posicionadas mais em cima e ao fundo.

O conceito de *itemset lattice* é usado para dividir as regras em sub-conjuntos. Desta forma, os relacionamentos entre sub-conjuntos vizinhos representam as relações de especialização e generalização existentes no *itemset lattice*, sendo possível navegar através desses relacionamentos. Assim, quando o usuário clica em uma regra, ele passa a visualizar todas as regras que representam especializações da regra clicada, onde por especialização entende-se a descida de um nível no *itemset lattice*. Da mesma forma, ao visualizar um conjunto de regras, o usuário pode querer fazer uma generalização, subindo um nível no *itemset lattice* e representando todas as regras que contenham



(a) Regra (b) Conjunto de regras

Figura 2.13: Abordagem proposta por Blanchard et al.

um item a menos que as regras exibidas. Como esta técnica apresenta apenas um sub-conjunto de regras de cada vez, ele escala bem para grandes volumes de regras. No entanto, o usuário não tem uma visão geral das regras, sendo necessário navegar entre vários sub-conjuntos até encontrar as regras mais interessantes (pode ser até que o usuário nem chegue a encontrar as regras mais interessantes). A utilização da posição para representar as métricas de interesse é uma vantagem, já que sabemos por Cleveland que a posição é a propriedade gráfica mais facilmente percebida pelo ser humano. Mas nesta proposta isso ajuda apenas a identificar as regras mais interessantes dentro do sub-conjunto de regras exibido em um dado instante, que pode não conter as regras mais interessantes do conjunto inteiro. Uma outra desvantagem desta técnica é que ela exige mais do que simples movimentos e cliques de mouse na interação, criando dificuldades para usuários pouco habituados a interações mais complexas.

Uma outra abordagem interessante é a proposta por Ong et al. (2000). Na sua proposta, as regras são organizadas em matriz, porém os eixos são usados para representar os valores de suporte e confiança das regras, onde cada regra é representada no ponto dado pelo seu suporte e confiança (Figura 2.14). Para representar as regras, são usadas formas geométricas, cujas propriedades gráficas podem ser usadas para representar outras métricas de interesse (além das já representadas nos eixos) ou para representar determinados itens.

As principais vantagens desta abordagem são: escalar bem tanto para grandes volumes de regras quanto para regras com muitos itens; permitir uma visão geral do conjunto de regras; utilizar a propriedade de posição no espaço para representar as métricas de interesse mais importantes; utilizar um paradigma bastante familiar, que

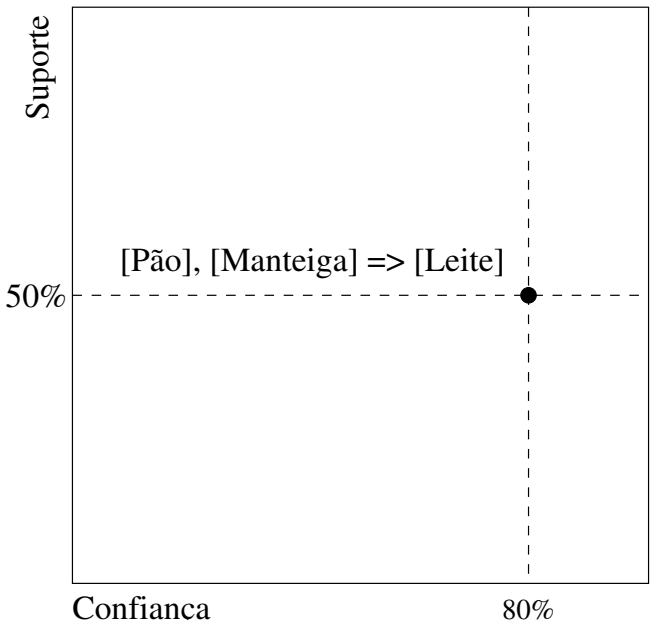


Figura 2.14: Matriz suporte×confiança.

lembra muito um *scatter-plot*; não exigir interações humanas muito complexas. A principal desvantagem é a sobreposição de pontos, já que várias regras podem apresentar valores iguais para as métricas representadas nos eixos. A Figura 2.15 mostra a implementação de Ong et al. para esta abordagem.

A maioria das técnicas apresentadas, tanto baseadas em matrizes quanto baseadas em grafos, oferecem a possibilidade de se filtrarem as regras pelos seus valores de suporte e confiança e pela presença de itens específicos – no antecedente ou no conseqüente – através de *sliders* e *check-boxes*. Além da técnica proposta por Kuntz et al. (2000), que se baseia inteiramente no conceito de *itemset lattice*, todas as outras abordagens podem se utilizar desse conceito para organizar as regras de forma hierárquica, como já o faz Blanchard et al.

2.3.2 Entendendo as Regras

O segundo problema de interação em sistemas de mineração de regras de associação está relacionado ao entendimento por parte do usuário dos conceitos relacionados à técnica, necessários para a utilização do sistema. Este problema não tem recebido a mesma atenção que o problema anterior: poucos trabalhos significativos foram realizados na tentativa de auxiliar o usuário a entender as regras de associação. Dentre eles, merece destaque o trabalho de Hofmann et al. (2000), no qual é proposta a utilização da técnica de *Mosaic plots* para auxiliar o usuário a compreender a importância de uma determinada regra de associação. (*Mosaic plots* são uma metáfora visual para as

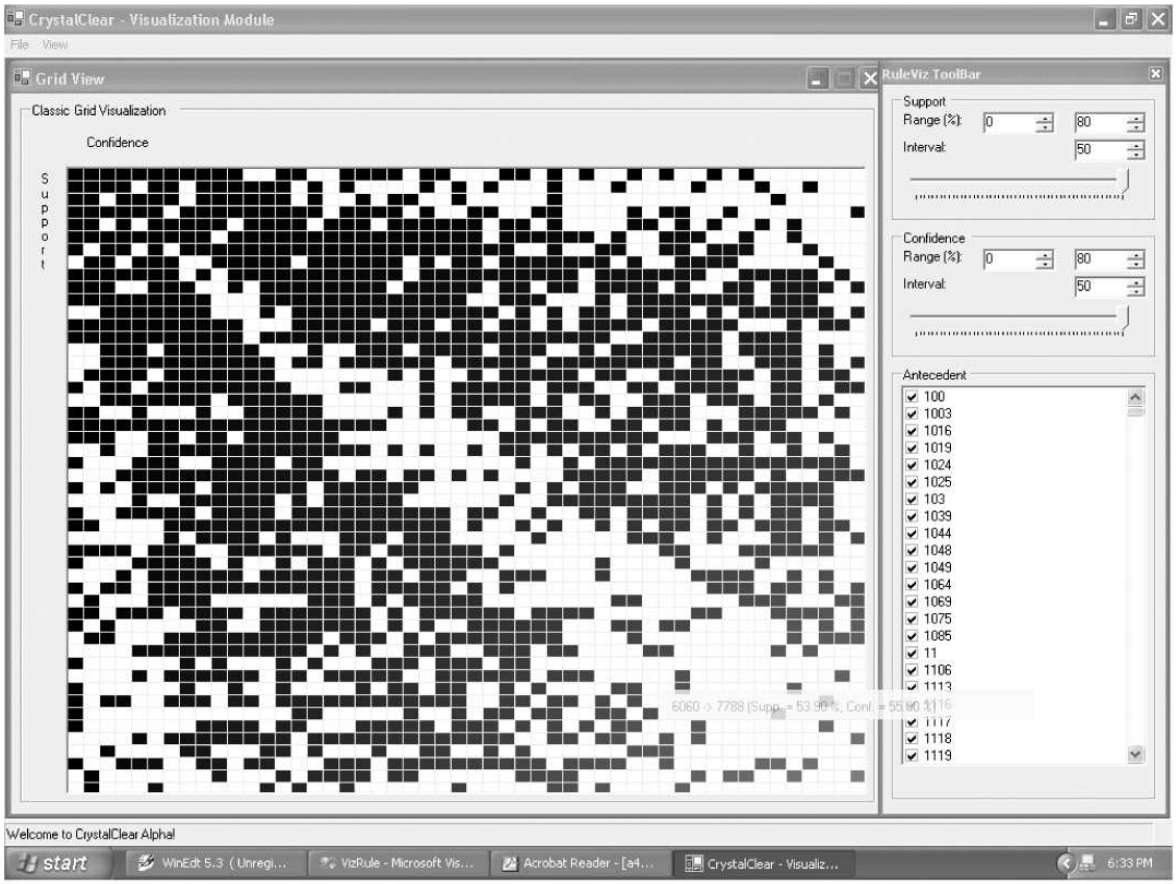


Figura 2.15: Abordagem proposta por Ong et al.

tabelas de contingência, conceito estatístico por trás de uma regra de associação.)

A Figura 2.16 representa várias regras de associação (tal como é proposto por Hofmann et al.), permitindo a comparação entre elas e facilitando o entendimento da importância relativa de cada uma delas. Na parte de baixo da figura, são representados os itens que podem aparecer no antecedente das regras (*heineken*, *coke* e *chicken*). A cor preta representa a presença do item e a cor branca representa a sua ausência. Na parte de cima são representados os conseqüentes das regras. No exemplo, o conseqüente pode ser *not sardines* ou *sardines*, ou seja, ausência ou presença do item *sardines*. As regras são representadas verticalmente, duas por coluna. Assim, as duas regras representadas na coluna mais à direita são:

$$[\text{chicken}],\ [\text{coke}],\ [\text{heineken}] \Rightarrow [\text{sardines}]$$
$$[\text{chicken}],\ [\text{coke}],\ [\text{heineken}] \Rightarrow [\text{not sardines}]$$

O suporte de ambas as regras é dado pela largura da coluna. A confiança da primeira regra é dada pela altura da parte preenchida da coluna (*sardines*) e a confiança da segunda regra é dada pela altura da parte não preenchida da coluna (*not sardines*).

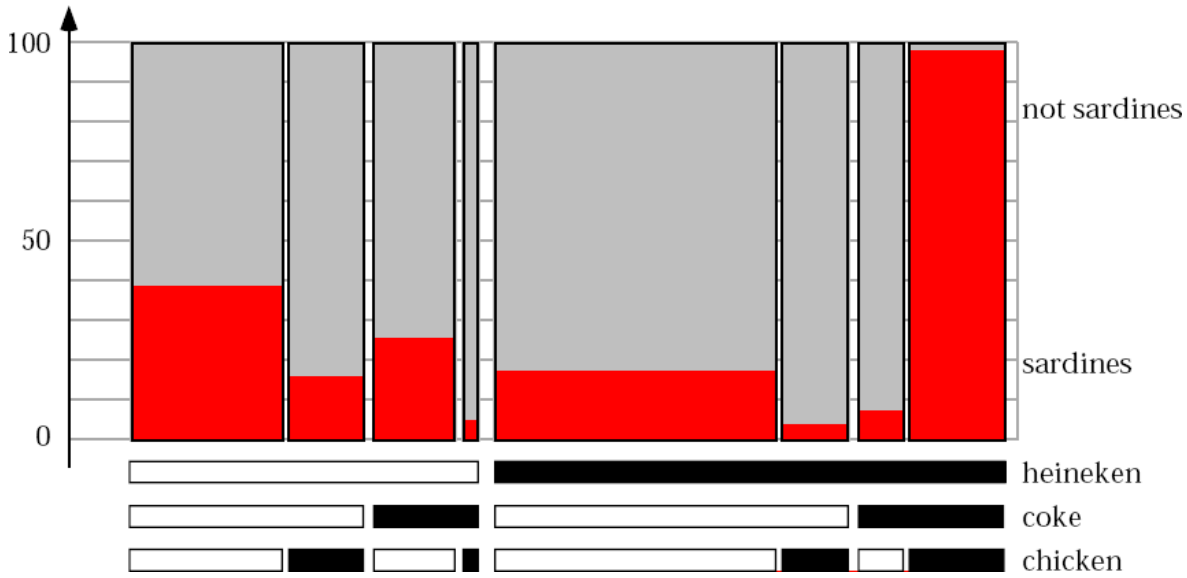


Figura 2.16: Abordagem de Hofmann et al. para a visualização e entendimento das regras. O suporte é dado pela área da coluna; a confiança é dada pelo percentual de preenchimento. A regra de maior confiança é a regra [chicken], [coke], [heineken] ⇒ [sardines].

A regra com maior confiança dentre as representadas é exatamente a regra [chicken], [coke], [heineken] ⇒ [sardines].

Esse trabalho, no entanto, apóia usuários que já possuem conhecimento sobre o que seja uma regra de associação e outros conceitos relevantes relacionados, mas não auxilia neste entendimento. O seu objetivo é oferecer aos usuários que já compreendem bem esses conceitos uma ferramenta para avaliar visualmente a importância de uma regra dentro do contexto das regras relacionadas a ela.

O nosso objetivo nesse aspecto é apoiar o projeto de sistemas de mineração de regras de associação para usuários que não têm conhecimento sobre as técnicas utilizadas e conceitos relacionados. Para isso, levantamos os aspectos relacionados ao desafio do projetista de comunicar a estes usuários os conceitos específicos necessários para que eles possam entender a solução proposta no sistema e utilizá-los com eficiência.

2.4 Sumário

Neste capítulo, fizemos uma revisão da bibliografia relacionada ao nosso trabalho. Como vimos, os estudos no campo da visualização da informação (tanto os relacionados à exibição estática da informação quanto os relacionados a aspectos interativos) podem contribuir para o desenvolvimento de interfaces de visualização de regras de associação de melhor qualidade. Vimos também que a mineração de dados já vem

sendo combinada com a visualização da informação de diferentes formas, na chamada mineração visual de dados.

Por fim, analisamos o que já foi feito especificamente com relação à interação em sistemas de mineração de regras de associação. Vimos que existem diversas abordagens para visualização das regras, com destaque para as abordagens baseadas em uma matriz antecedente×conseqüente e as baseadas em grafos. Esses trabalhos tratam essencialmente do problema de se encontrar as regras mais interessantes em meio a grandes volumes de regras. Quanto ao problema de entendimento das regras, vimos que são poucos os trabalhos existentes.

Capítulo 3

Estratégias de Visualização de Regras de Associação

Neste capítulo, apresentamos as estratégias para visualização de regras de associação que desenvolvemos neste trabalho. Embora o problema de visualização de regras de associação já tenha sido abordado de diferentes formas, como vimos no capítulo anterior, não existe nenhum estudo que traga uma avaliação dessas abordagens com usuários reais. As estratégias apresentadas aqui serviram para a avaliação de duas diferentes abordagens, além de subsidiar o levantamento dos desafios de entendimento que constituem o segundo problema de interação em mineração de regras de associação. Do ponto de vista prático, essas interfaces têm sido muito úteis, já que vêm sendo utilizadas por usuários reais.

3.1 Visualização Estrutural

A primeira estratégia que desenvolvemos é o que chamamos de Visualização Estrutural. Devemos nos recordar que uma regra é formada por um antecedente e um conseqüente. Além disso, existe um conceito relacionado a regras de associação chamado de *itemset lattice*, que define uma hierarquia entre as regras. Esses conceitos (antecedente, conseqüente e *itemset lattice*) constituem a estrutura de uma regra e a estrutura do conjunto de regras. Daí o nome Visualização Estrutural: a visualização proposta nesta primeira estratégia expõe essas estruturas para o usuário.

Esta versão utiliza uma abordagem baseada numa matriz antecedente×conseqüente (Figura 3.1). Como vimos no Capítulo 2, nesta abordagem os *itemsets* são distribuídos ao longo dos eixos e as regras são representadas por figuras geométricas posicionadas na interseção da linha correspondente ao antecedente com a coluna correspondente ao conseqüente da regra. Na nossa implementação, os *itemsets* são dispostos nos eixos

na ordem decrescente dos seus suportes. Assim, as regras com maior suporte dentre as regras exibidas aparecem geralmente no canto superior esquerdo da tela. Como o espaço na tela é limitado, é necessária a utilização de barras de rolagem em ambos os eixos.

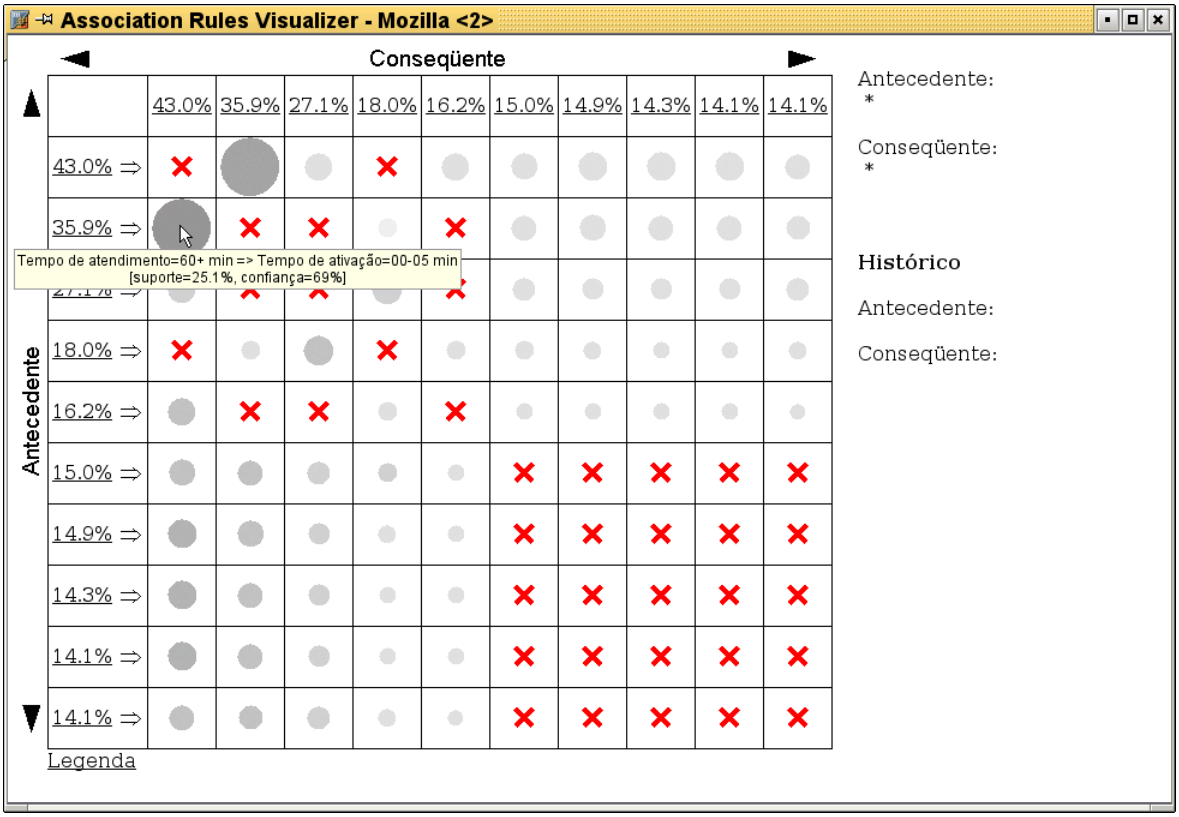


Figura 3.1: Tela inicial da primeira versão. Área e tom de cinza do círculo que representa uma regra dá uma medida do quanto ela é interessante.

As regras são representadas por círculos, e o conteúdo de uma regra é exibido sempre que se posiciona o cursor do *mouse* em cima do círculo correspondente à regra. A área do círculo é usada para representar o suporte, e a confiança é representada pela intensidade do tom de cinza (Figura 3.2). Para otimizar a utilização das diversas combinações de tamanho e tom de cinza, a escala utilizada é sempre ajustada de acordo com os valores máximos de suporte e confiança encontrados no conjunto de regras. Como podemos notar na Figura 3.2 por exemplo, a escala do suporte vai de 0 a 25,1%, o que demonstra que a regra com maior suporte no conjunto de regras em questão apresenta um suporte igual a 25,1%. Sem o ajuste na escala, apenas os três primeiros círculos seriam usados neste caso, dificultando a comparação das regras.

Uma outra característica importante dessa versão é que ela utiliza o conceito de *itemset lattice* para organizar as regras de forma hierárquica. Assim, na tela inicial são mostrados apenas os *itemsets* de tamanho 1, tanto no antecedente quanto no con-

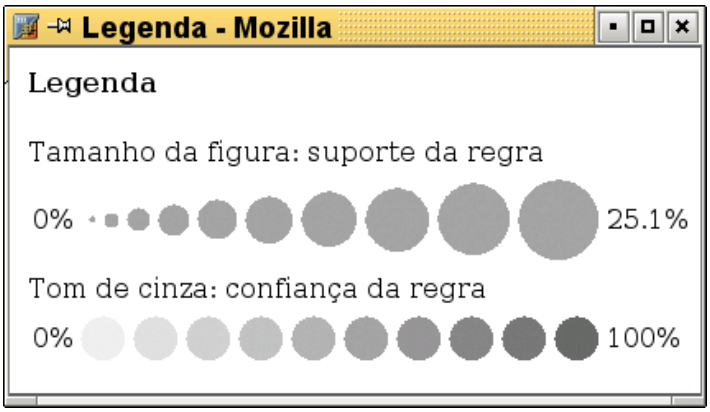


Figura 3.2: Legenda usada na primeira versão.

seqüente, e as regras exibidas inicialmente são todas regras com apenas um item no antecedente e um item no conseqüente (e.g., [Pão]⇒ [Leite]). Para visualizar regras com mais itens, é necessário clicar em um *itemset* no eixo vertical ou no eixo horizontal, e esse eixo passará a exibir todos os *itemsets* que contenham o *itemset* clicado e mais um item (o que corresponde a uma operação de especialização, ou uma descida de nível no *itemset lattice*).

3.2 Visualização por Métricas de Interesse

A segunda abordagem que desenvolvemos destaca as métricas de interesse em detrimento da estrutura das regras. O foco aqui é na identificação das regras mais interessantes, antes mesmo da compreensão do significado das regras. Para desenvolver esta versão, nos baseamos no trabalho de Ong et al. No entanto, estendemos o seu trabalho nos seguintes sentidos:

- Apresentamos uma implementação desta abordagem que atende a requisitos específicos, dentre eles: volumes de regras na casa das centenas de milhares, armazenamento das regras em uma máquina e exibição em outra (arquitetura distribuída), perfis específicos de usuários;
- Comparamos empiricamente esta abordagem com outras abordagens;
- Avaliamos esta abordagem com usuários reais, em aplicações do mundo real;
- Adicionamos novas funcionalidades, como a possibilidade de substituir as métricas nos eixos por três outras métricas (*lift*, *leverage* e convicção), função de *drill-down* para visualizar os dados atômicos que são sumarizados por uma regra, opções de filtro mais avançadas, maior detalhamento das regras selecionadas, dentre outras;

- Habilitamos a aplicação para ser utilizada via Web, o que permitiu uma adoção mais ampla do sistema, além de contribuir com estudos de caracterização a partir da geração de *logs*.

Esta versão organiza o conjunto de regras em uma matriz suporte×confiança (Figura 3.3). Na sua configuração padrão, o eixo vertical representa o suporte e o eixo horizontal representa a confiança. As regras são representadas por pequenos quadrados azuis, na posição dada pelos seus valores de suporte e confiança. Uma das principais vantagens dessa abordagem é a facilidade com que são identificadas as regras de maior suporte e/ou confiança. Como vimos no Capítulo 2, Cleveland demonstrou empiricamente a maior precisão da propriedade de posição para representar informações quantitativas.

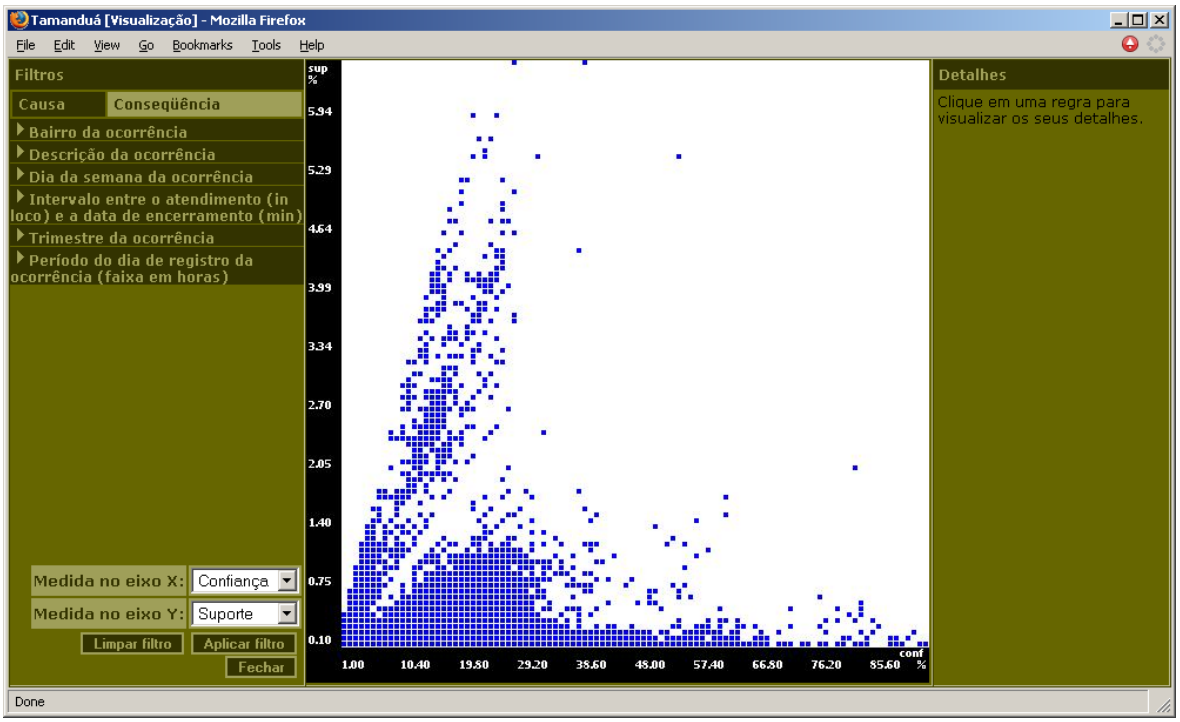


Figura 3.3: Tela inicial da segunda versão. A posição da regra no gráfico dá uma medida do quanto ela é interessante.

Shneiderman (1996) sugere como um ponto de partida útil para o desenho de interfaces gráficas avançadas aquilo que ele chama de "Mantra da Busca Visual de Informações":

> Visão geral primeiro, *zoom* e filtro, e então detalhes sob demanda.

Este princípio guiou o desenvolvimento desta segunda versão, uma vez que o principal problema com a versão anterior era exibir para o usuário um sub-conjunto das regras

geradas sem dar a ele antes uma visão geral do conjunto inteiro de regras. Quando o sistema exibe inicialmente apenas um sub-conjunto da coleção, ele está realizando pelo usuário um *zoom* ou um filtro, sem que o usuário possa dizer se ele realmente deseja isso. Desta forma, o usuário não apenas tem dificuldade em entender a decisão tomada pelo sistema, como se aborrece por não se sentir no controle.

Após ter uma visão geral, o usuário pode detalhar as regras que ele desejar, por exemplo aquelas que apresentarem um maior suporte ou uma maior confiança (ou ambos). Basta clicar em um quadrado e as informações da regra são exibidas no painel de detalhes à direita (Figura 3.4). Em geral, quando o usuário está fazendo uma análise exploratória, esse é o comportamento esperado.

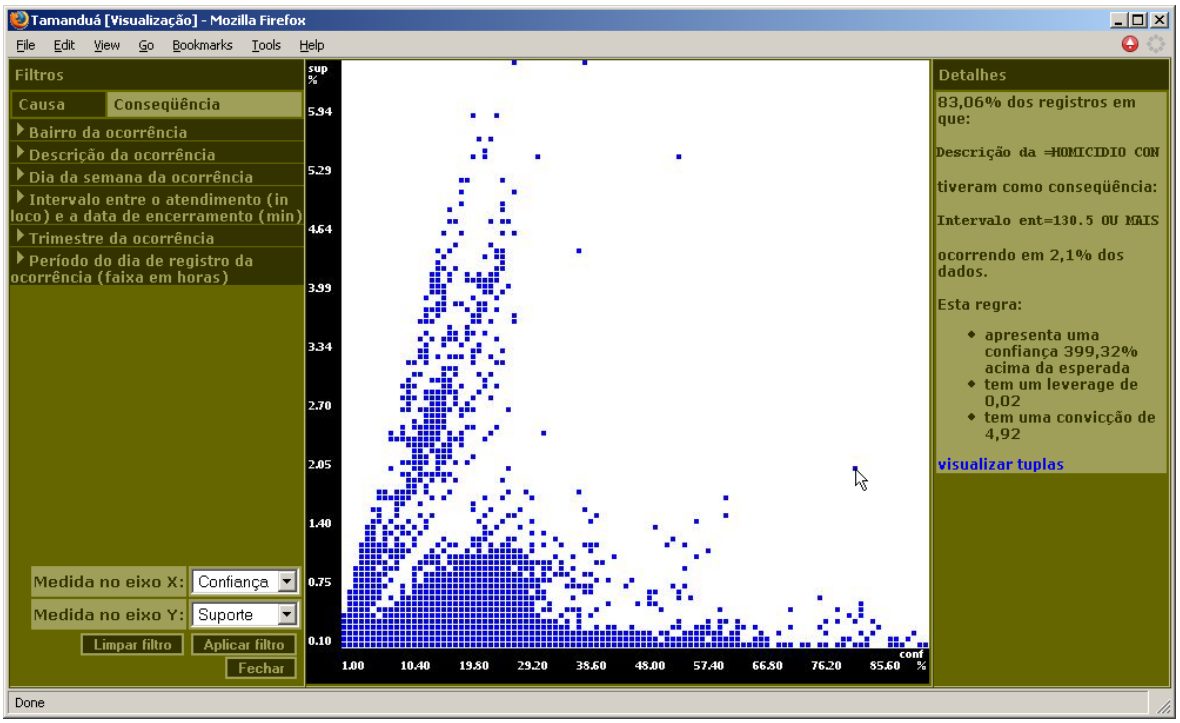
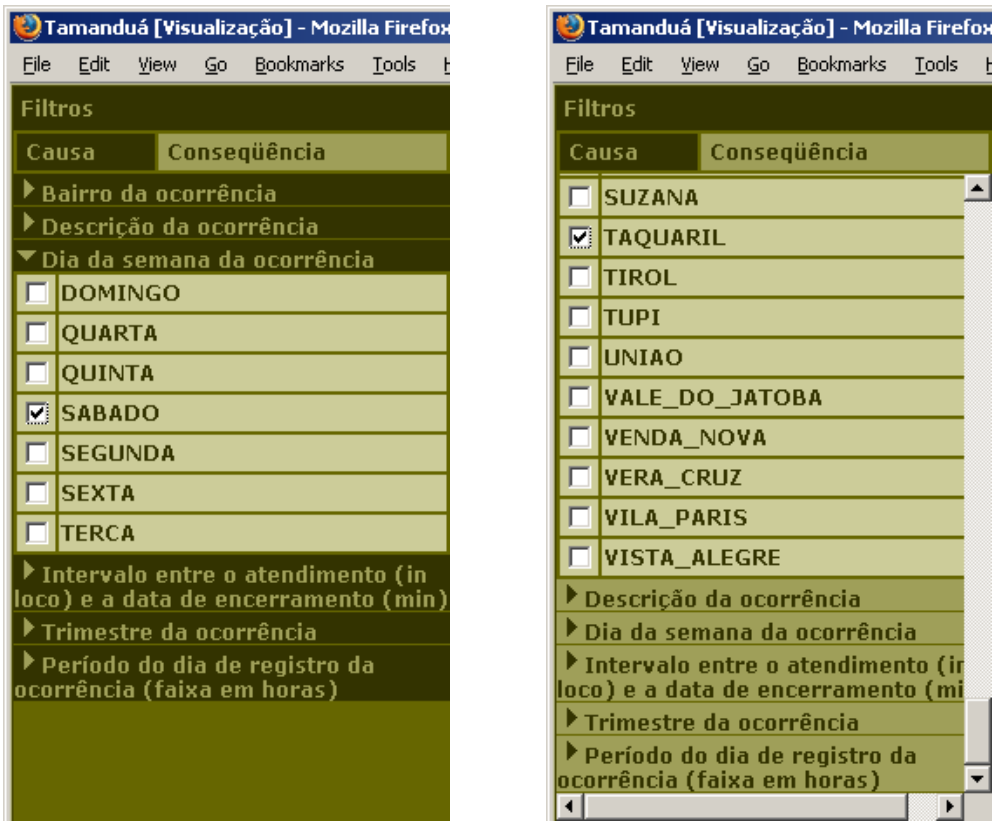


Figura 3.4: Painel de detalhes, que aparece quando o usuário clica em uma regra.

Por outro lado, se o usuário quiser testar uma hipótese específica ou fazer uma análise exploratória mais direcionada, ele irá desejar focar a sua busca. Para isso, ele pode se utilizar do mecanismo de filtro disponível no sistema. O mecanismo de filtro permite que o usuário encontre as regras que contenham itens específicos, podendo dizer se ele quer visualizar as regras que contenham determinados itens no antecedente ou no conseqüente (Figura 3.5). No exemplo, estamos filtrando todas as regras que tenha no antecedente o item *Dia=SÁBADO* e no conseqüente o item *Bairro=TAQUARIL*. O conjunto de regras que atendem a esse critério é bem menor, facilitando a análise (Figura 3.6).



(a) Filtrar as regras pela presença do item *Dia=SÁBADO* no antecedente.

(b) Filtrar as regras pela presença do item *Bairro=TAQUARIL* no conseqüente.

Figura 3.5: Mecanismo de filtro da segunda versão.

Esta versão permite também que o usuário altere as métricas de interesse representadas nos eixos (Figura 3.7). O suporte e a confiança, embora sejam as métricas mais usadas, quase nunca são as mais indicadas para se avaliar o interesse de uma regra. No nosso sistema, o usuário tem a opção de alterar as métricas de interesse nos eixos por três outras métricas: *lift*, *leverage* e convicção. Essas métricas são muito mais eficazes que o suporte e a confiança, e embora as três sejam bem parecidas entre si, as regras que são consideradas interessantes por uma raramente são as mesmas que são consideradas interessantes pelas outras.

O *lift* é calculado dividindo-se a confiança da regra pela freqüência do seu conseqüente. Assim, o *lift* dá uma medida do quanto a ocorrência do antecedente aumenta ou diminui a chance de ocorrência do conseqüente ou, em outras palavras, o quanto a confiança da regra é maior ou menor do que o esperado. Quando a confiança de uma regra é próxima do esperado, o *lift* assume um valor próximo de 1. À medida em que a confiança vai se afastando do valor esperado o valor do *lift* vai se afastando de 1. Em

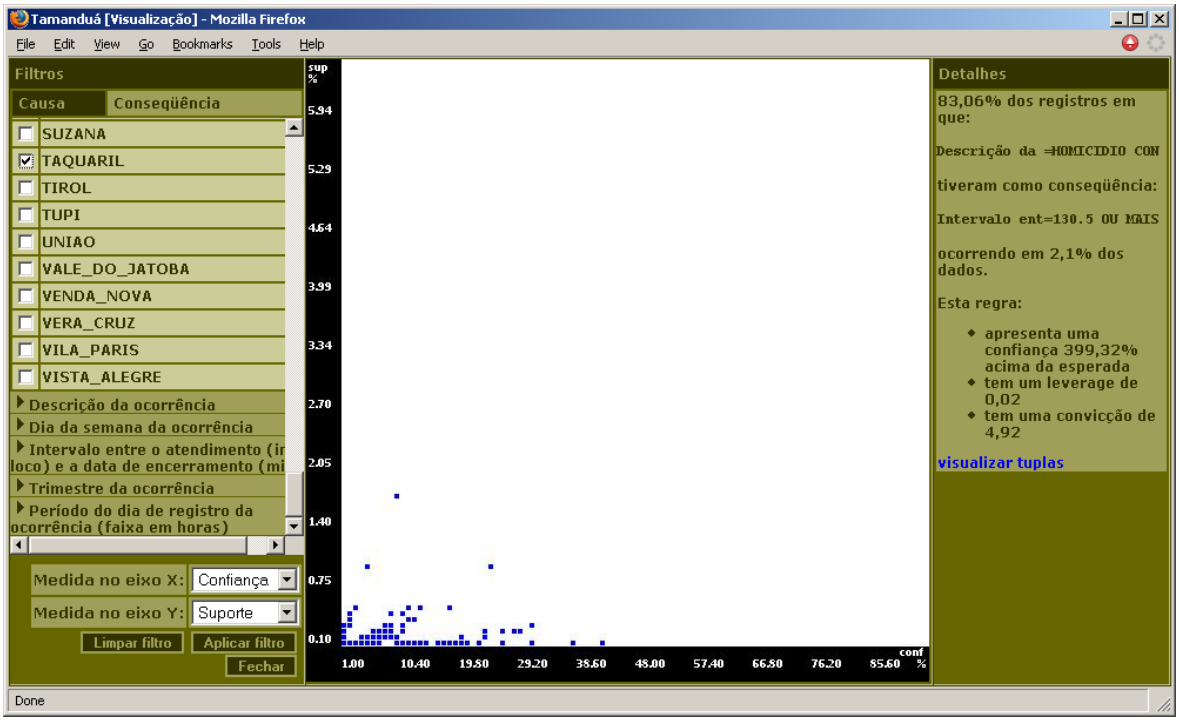


Figura 3.6: Resultado da aplicação do filtro.

Figura 3.7: Escolha das medidas de interesse que deverão ser representadas nos eixos.

geral, as regras que apresentam confiança baixa mas um *lift* alto são mais interessantes do que as regras que apresentam uma confiança alta e um *lift* próximo de 1, pelo fato de serem mais surpreendentes.

O *leverage* corresponde à diferença entre o suporte de uma regra e valor que seria esperado para esse suporte. Para calcular o "suporte esperado" de uma regra, multiplica-se a freqüência do antecedente pela freqüência do conseqüente. Assim, se o evento representado no antecedente de uma regra tiver uma freqüência alta na base, o mesmo ocorrendo com o evento representado no conseqüente, a tendência é que o suporte da regra seja alto. O *leverage* ajuda a avaliar o quanto o suporte de uma determinada regra é surpreendente, ou seja, diferente do esperado.

A convicção tem uma função semelhante à do *lift*, mas ela é calculada de forma indireta: em vez de se calcular o quanto a ocorrência do antecedente de uma regra aumentou a chance de ocorrência do conseqüente, calcula-se o quanto a ocorrência do antecedente diminuiu a chance da não ocorrência do conseqüente.

Uma outra funcionalidade do sistema é que ele permite que o usuário faça *drill-down* em uma regra para visualizar os dados que são sumarizados por ela. Vamos considerar novamente a nossa regra de exemplo:

$$[\text{Pão}], [\text{Manteiga}] \Rightarrow [\text{Leite}] \ (80.00, 50.00)$$

Se o usuário fizer *drill-down* nesta regra, ele irá visualizar todas as compras da padaria que incluíram Pão, Manteiga e Leite. Essas compras correspondem a 50% de todas as compras realizadas na padaria, que é o que significa o suporte da regra. O usuário pode querer fazer isso se ele quiser saber por exemplo quais foram os clientes que realizaram essas compras, quando elas foram realizadas etc.

Para esta estratégia, foi desenvolvido ainda um sistema de ajuda, onde são explicados os conceitos básicos de mineração de regras de associação e onde são fornecidos exemplos de utilização do sistema. Vale ressaltar que várias das funcionalidades implementadas para a segunda estratégia são independentes da estratégia em si, e poderiam ter sido implementados também na primeira estratégia. São elas: o painel de detalhamento de regras, as métricas de interesse adicionais, a funcionalidade de *drill-down*, além do próprios sistema de ajuda. É importante fazer essa distinção para evitar que a comparação entre as duas estratégias seja tendenciosa.

3.3 Sumário

Vimos assim duas abordagens completamente diferentes para visualização de regras de associação. A primeira, Visualização Estrutural, dá ênfase à estrutura tanto de uma regra quanto do conjunto de regras. Ela organiza as regras em uma matriz antecedente×conseqüente, e utiliza ainda o conceito de *itemset lattice* para apresentar as regras de forma hierárquica. A segunda abordagem privilegia as métricas de interesse de uma regra. As regras são organizadas em uma matriz suporte×confiança, sendo que essas métricas podem ser substituídas por até outras três métricas distintas: *lift*, *leverage* e convicção. No capítulo seguinte, apresentamos os resultados da avaliação dessas interfaces sob diferentes aspectos.

Capítulo 4

Avaliações

Neste capítulo, apresentamos os resultados de estudos sobre a utilização da nossa interface por usuários reais. Esses estudos foram realizados durante os cursos de treinamento do sistema Tamanduá, do qual a nossa interface é componente. Os principais objetivos desses estudos eram (1) avaliar o quanto os usuários se sentem confortáveis com as estratégias de visualização desenvolvidas, (2) avaliar a usabilidade das nossas implementações, (3) levantar os aspectos relacionados à mineração de regras de associação que são mais difíceis de serem compreendidos pelos usuários e avaliar o custo associado à não compreensão desses aspectos e (4) caracterizar o comportamento geral dos usuários de sistemas de mineração de regras de associação.

A seguir, apresentamos o sistema Tamanduá e mostramos o contexto no qual as nossas experiências se realizaram. Logo depois, apresentamos os principais resultados dessas experiências.

4.1 O Sistema Tamanduá

No Departamento de Ciência da Computação da UFMG foi desenvolvido um sistema de mineração de dados, o Tamanduá, que possui um módulo de mineração de regras de associação. Entre outras coisas, o Tamanduá tem o objetivo de apoiar a gestão e decisão governamentais, em particular em tarefas de auditoria relacionadas a compras e contratações, e também como ferramenta de análise para cientistas sociais, em pesquisas na área de criminalidade e segurança pública. A grande maioria dos usuários do Tamanduá, embora especialista no seu domínio de aplicação, é leiga em mineração de dados, o que motivou o desenvolvimento de uma interface que pudesse propiciar a esses usuários uma melhor utilização do sistema. O Tamanduá já vem sendo utilizado em iniciativas piloto em diversas instituições públicas brasileiras, incluindo:

- Auditoria Geral do Estado de Minas Gerais (AUGE);

- Centro de Estudos em Criminalidade e Segurança Pública da UFMG (CRISP);
- Hospital das Clínicas da UFMG (HC/UFMG);
- Ministério da Saúde;
- Secretaria de Estado de Planejamento e Gestão de Minas Gerais (SEPLAG), e
- Secretaria de Logística e Tecnologia da Informação do Ministério do Planejamento, Orçamento e Gestão (SLTI/MP).

Além dessas instituições, o sistema Tamanduá vem sendo utilizado também em cursos de mineração de dados nas seguintes instituições:

- Companhia de Tecnologia da Informação do Estado de Minas Gerais (PRODEMGE);
- Fundação João Pinheiro, e
- Universidade Federal de Minas Gerais.

O sistema Tamanduá foi desenvolvido dentro do paradigma de computação orientada a serviços. Isto significa que vários módulos do sistema são implementados como serviços independentes que são conjugados, em tempo de projeto ou de execução, para a realização de uma determinada tarefa (Figura 4.1).

Uma tarefa típica de mineração de dados é executada no Tamanduá através da atuação conjunta de cinco servidores: o primeiro servidor é o servidor de aplicação, que é responsável por oferecer ao usuário uma interface gráfica para definição da tarefa de mineração que ele quer realizar; o segundo servidor é o servidor de mineração, que é responsável por buscar os dados a serem minerados e disparar a execução dos algoritmos sobre esses dados; o terceiro servidor é o servidor de dados, que é o repositório de onde o servidor de mineração busca os dados a serem minerados; o quarto servidor é o servidor de processamento, responsável pela execução dos algoritmos sobre os dados; e o quinto servidor é o servidor de visualização, que é o repositório onde são armazenados os resultados da mineração. A integração entre os servidores é feita através de *Web Services*, um padrão para chamada remota de procedimentos implementado a partir de tecnologias abertas, como HTTP, XML e outras.

Um dos principais módulos do sistema Tamanduá é o módulo de mineração de regras de associação. Os resultados de uma tarefa de mineração de regras de associação podem chegar facilmente a tamanhos próximos de 100 MB. Para que o usuário possa visualizar os resultados da mineração, é necessário que o servidor de aplicação acesse o

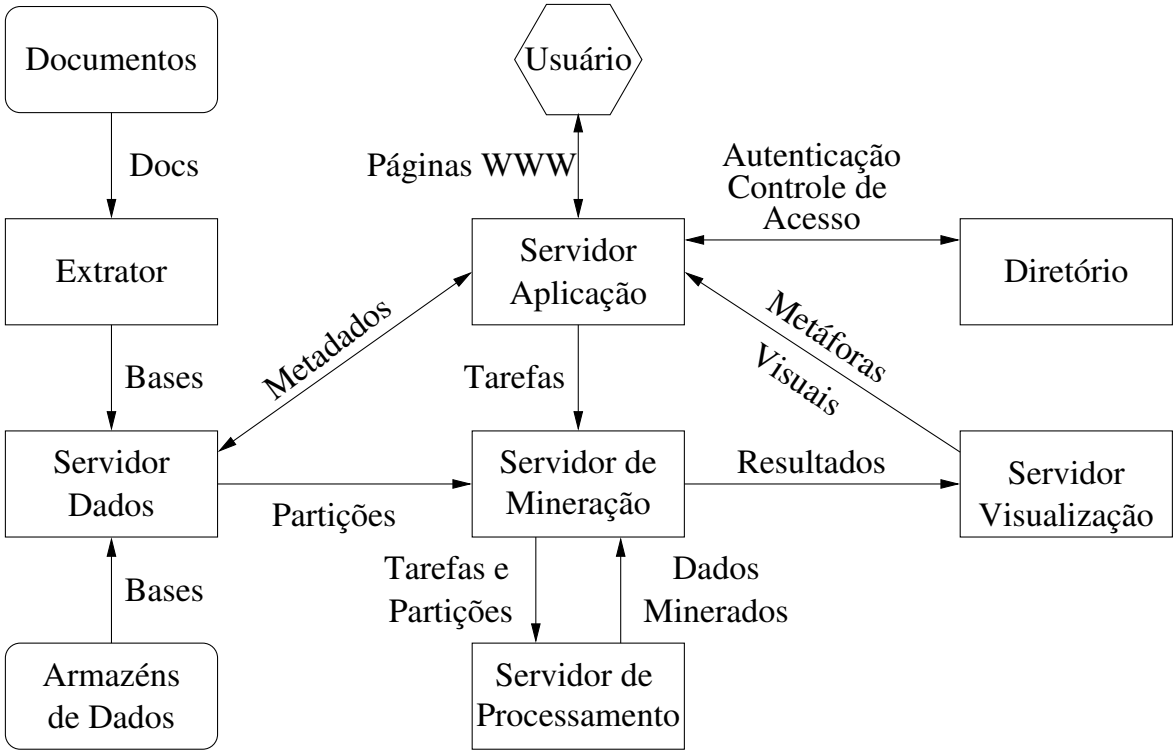


Figura 4.1: Arquitetura do sistema Tamanduá.

servidor de visualização onde estão armazenados os resultados e requisite as visualizações necessárias, que são exibidas ao usuário como páginas Web, geradas pelo servidor de aplicação.

Deste cenário, surgem três importantes restrições para o projeto de uma interface de visualização de regras de associação para o Tamanduá:

1. Em função do tamanho, os resultados da mineração não podem ser copiados para a máquina do usuário, ou pelo menos não deve-se assumir que sempre será viável copiá-los;

2. A interface com o usuário deve ser acessível através de um navegador Web, sem a necessidade de instalação de nenhum *plug-in*, e

3. Os resultados a serem visualizados devem ser enviados do servidor de visualização onde estão armazenados para o cliente (que no caso é o servidor de aplicação) através de uma interface de *Web Services*.

4.1.1 Treinamentos

Todos os usuários do Tamanduá, antes de começarem a utilizar o sistema, passam por um treinamento de cerca de 4 horas de duração, quando são explicados os conceitos

básicos de mineração de dados e mineração de regras de associação, e apresentadas as funcionalidades principais do sistema. O curso tem uma carga horária prática de aproximadamente 2 horas, quando os usuários têm a oportunidade de utilizar o sistema pela primeira vez. Os resultados que apresentamos neste capítulo foram obtidos a partir da observação do comportamento dos usuários durante as práticas realizadas nesses cursos.

Ao final do treinamento, é solicitado aos usuários que respondam a um questionário de avaliação do sistema, baseado no QUIS (*Questionnaire for User Interaction Satisfaction*) (Chin et al., 1988). Vamos apresentar também os resultados dessa avaliação. Além disso, analisamos os *logs* gerados com a utilização do sistema tanto durante os cursos quanto em atividades extra-classe (uma vez que os usuários têm acesso ao sistema após a realização dos cursos). Os resultados das análises desses *logs* também serão reportados aqui.

4.1.2 Usuários

A maioria dos usuários do Tamanduá incluídos na nossa análise possuem curso superior completo, sendo que alguns deles possuem ou estão cursando algum tipo de pós-graduação. Muitos possuem certa experiência com técnicas básicas de análise estatística, mas o conhecimento da área de mineração de regras de associação se resume ao obtido no treinamento. Embora alguns deles sejam oriundos da área de tecnologia da informação (TI), a maioria absoluta tem formação em áreas não relacionadas à área de TI, como ciências sociais, economia, administração, medicina etc. Acreditamos que a grande maioria dos usuários potenciais de mineração de dados apresente um perfil semelhante. No Apêndice A incluímos o questionário utilizado no levantamento do perfil dos usuários do Tamanduá.

As nossas experiências foram realizadas durante 5 cursos, oferecidos para 5 diferentes instituições, em um total de 92 usuários. A Tabela 4.1 mostra as instituições que participaram desses cursos, e a quantidade de participantes por instituição.

Instituição	Usuários	Homens	Mulheres
AUGE	10	7	3
CRISP	11	6	5
HC/UFMG	11	6	5
MP/SLTI	22	16	6
PRODEMGE	40	26	14

Tabela 4.1: Cursos de Treinamento no Sistema Tamanduá, nos quais foram realizadas as nossas experiências.

4.1.3 Tarefas

O Tamanduá vem sendo utilizado até o momento para suportar duas tarefas básicas. A primeira delas é a análise exploratória de dados. O papel da mineração de regras de associação neste caso é auxiliar no levantamento inicial de hipóteses, que serão melhor analisadas posteriormente através de outras ferramentas, principalmente estatísticas. A segunda tarefa consiste na identificação de informações a partir das quais possa ser tomada uma medida imediata ou programada uma ação específica. É o que acontece por exemplo em tarefas de auditoria, quando o sistema é usado para ajudar os auditores a identificarem situações suspeitas a partir das quais são disparados processos de investigação.

Durante os cursos, os usuários usam livremente o sistema para realizar minerações nas bases de dados com as quais eles estão acostumados a trabalhar no seu dia-a-dia. Em apenas um dos cursos foi fornecido um roteiro com as tarefas que deveriam ser realizadas pelos usuários durante o curso. Este roteiro incluía questões envolvendo análise exploratória, teste de hipóteses e refinamento de hipóteses. Por exemplo, as seguintes tarefas faziam parte desse roteiro:

- Encontre as três regras que você julgar mais interessantes;
- Verifique a seguinte hipótese: "O índice de homicídios e tentativas de homicídios nas favelas é maior que a média geral";
- Supondo que já se saiba que uma ocorrência aconteceu no centro da cidade, às 9 horas da manhã de um domingo, no mês de dezembro, qual a informação que pode ser deduzida com maior chance de acerto?

4.2 Resultados

Existem três dimensões nas quais podemos avaliar a nossa interface. A primeira delas diz respeito à visualização em si, ou seja, à metáfora visual utilizada para representar as regras de associação e a adequação desta metáfora, tanto ao tipo de dados que estamos querendo visualizar (as regras) quanto às tarefas cognitivas associadas a esses dados. Neste aspecto, mostramos ao longo do texto as vantagens e desvantagens das duas estratégias que desenvolvemos, com base na literatura de visualização da informação e de mineração de regras de associação. Nas nossas experiências com os usuários do sistema Tamanduá, avaliamos empiricamente essas estratégias, para o perfil específico dos nossos usuários.

Um dos principais resultados dessas avaliações é que pudemos obter fortes indícios da vantagem da abordagem baseada em matriz suporte×confiança (Visualização por

Métricas de Interesse) sobre a abordagem baseada em matriz antecedente×conseqüente (Visualização Estrutural). Esta constatação pôde ser feita a partir da facilidade encontrada pelos usuários na utilização da primeira comparada com a dificuldade encontrada por eles na utilização da segunda. A estratégia de Visualização Estrutural foi avaliada informalmente com 5 usuários, todos usuários com conhecimentos prévios de mineração de dados ou de técnicas estatísticas. As avaliações revelaram que a solução de *design* proposta nesta versão era de difícil compreensão, mesmo para esses usuários mais sofisticados. Os usuários tinham muita dificuldade em decidir para onde navegar a partir da tela inicial, o que se justificava basicamente por dois motivos: em primeiro lugar a navegação requeria que o usuário tivesse uma hipótese inicial, o que nem sempre era o caso; depois, a navegação exigia do usuário um conhecimento do conceito de *itemset lattice*, já que clicar em um *itemset* conceitualmente representava descer um nível no *itemset lattice*. A dificuldade encontrada por esses usuários desencorajou o investimento em novas versões desta estratégia.

Por outro lado, as nossas experiências revelaram a facilidade de uso da estratégia de Visualização por Métricas de Interesse. A maioria dos usuários que realizou o curso não teve problema com a visualização. Dos 14 usuários que responderam ao questionário, 9 mencionaram nas questões abertas (Figura 4.2) a facilidade de uso do sistema.

Além das experiências realizadas no curso, que podem ser consideradas avaliações informais, realizamos ainda uma avaliação formal com 3 usuários. O Apêndice B traz a lista de tarefas utilizada nessa avaliação. Como o comportamento dos usuários no domínio de mineração de regras de associação nunca foi estudado, optamos por uma análise qualitativa. Além disso, a mineração de dados constitui um processo de obtenção do conhecimento, e as métricas padrões de usabilidade, como eficiência e quantidade de erros, não parecem adequadas. Os resultados dessa avaliação formal serviram apenas para confirmar os resultados obtidos nas avaliações informais, e não apresentaram nenhum dado novo.

A segunda dimensão diz respeito aos princípios gerais de *design*, como consistência, *feedback*, prevenção de erros etc. (Norman, 1988). A seguir analisamos alguns dos resultados relacionados a este aspecto. Embora este aspecto seja importante para o sistema Tamanduá do ponto de vista prático, ele não é muito relevante para os objetivos deste trabalho, pois não representa um desafio do ponto de vista de pesquisa.

As avaliações revelaram problemas comuns de usabilidade, como:

- Inconsistência: algumas vezes são usados termos diferentes para representar a mesma coisa; outras vezes ações parecidas produzem resultados diferentes;
- Pouco *feedback*: o sistema falha em manter o usuário informado da sua situação atual; algumas ações não geram nenhum resultado, e o usuário fica sem saber o

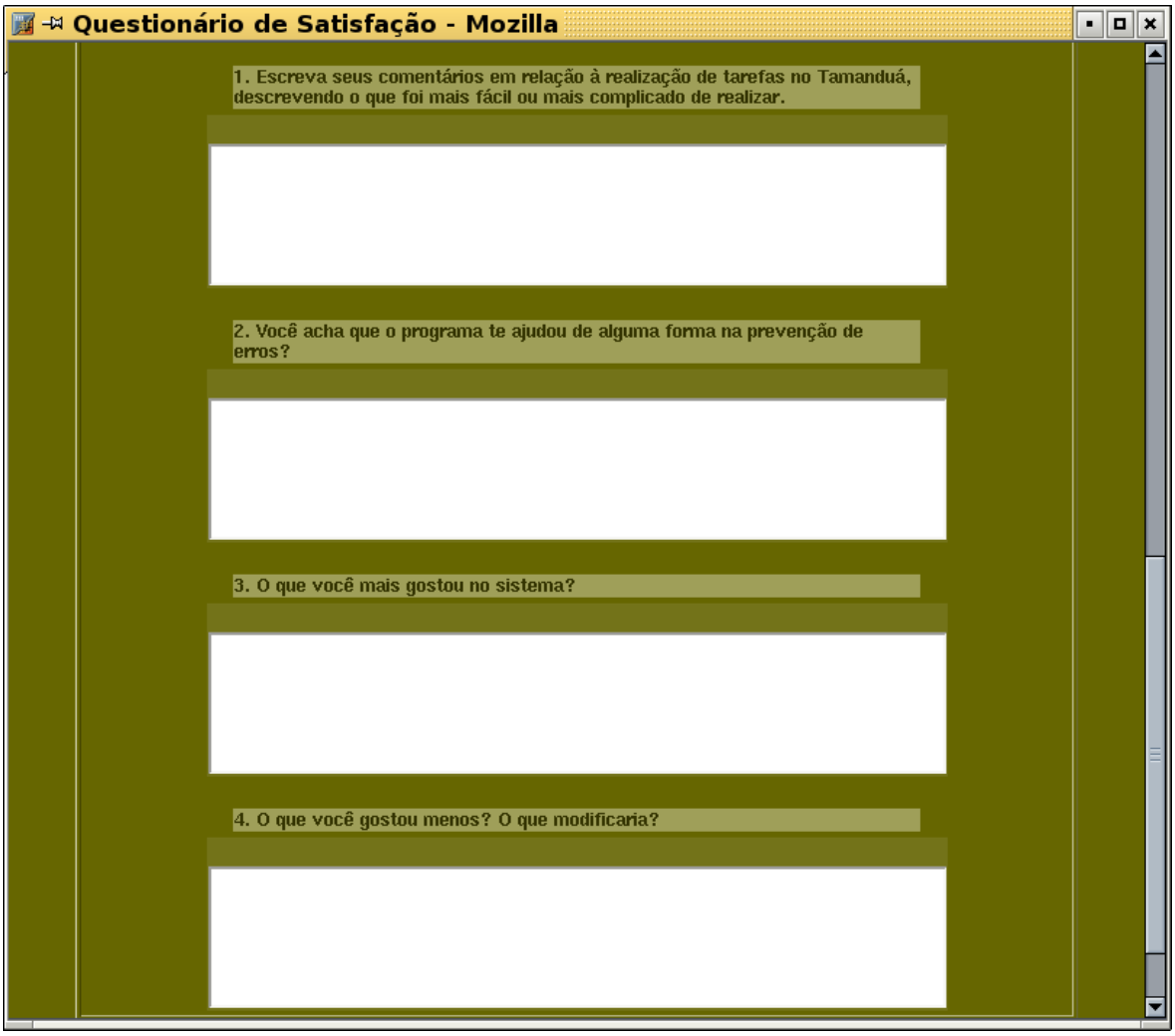


Figura 4.2: Questionário de satisfação: questões abertas.

que aconteceu;

- Poucos mecanismos de prevenção de erros, e
- Ausência de suporte a ações de desfazer e refazer.

Esses e outros problemas de usabilidade poderiam ser identificados através de avaliações por inspeção, quando especialistas em usabilidade revisam uma interface para determinar a sua conformidade com uma lista de heurísticas qualquer, como as dez heurísticas de usabilidade (Nielsen, 2005) ou as oito regras de ouro do desenho de interfaces (Shneiderman e Plaisant, 2004).

Nas questões fechadas do questionário (Figura 4.3), foram obtidos resultados interessantes relacionados esses atributos. A Tabela 4.2 mostra a média das notas de cada uma das questões, numa escala de 1 a 9, além do desvio padrão associado a cada uma delas.

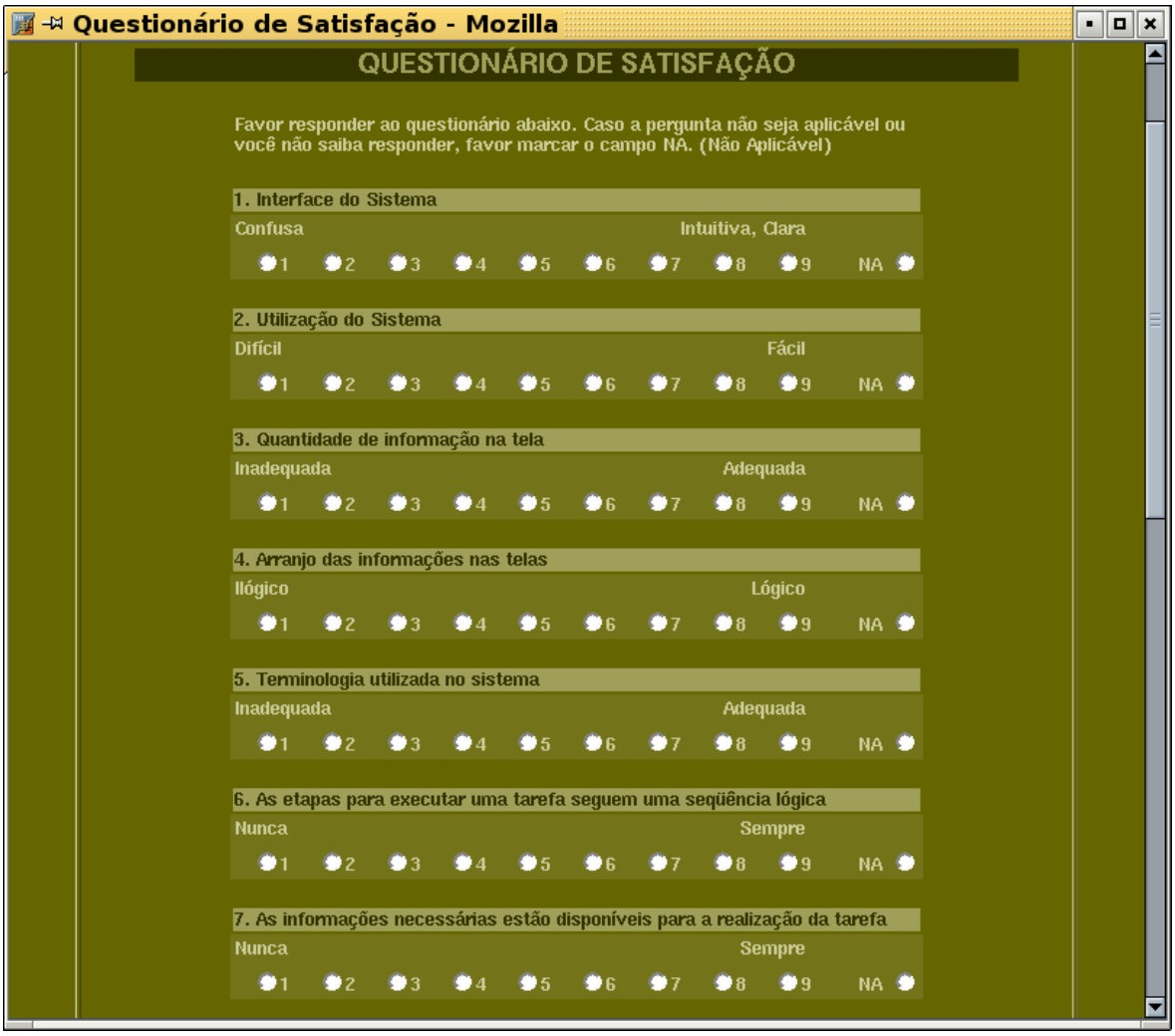


Figura 4.3: Questionário de satisfação: questões fechadas.

A terceira dimensão diz respeito ao entendimento, por parte dos usuários, dos conceitos específicos de mineração de regras de associação. Como pudemos constatar nas nossas experiências, este é o principal problema enfrentado pelos usuários na sua interação com o sistema, como Hofmann et al. já havia alertado. 7 usuários dos que responderam ao questionário reclamaram da dificuldade de entendimento dos conceitos utilizados no sistema. Interpretamos esses resultados da seguinte maneira: a metáfora visual utilizada é intuitiva e os usuários conseguem utilizá-la satisfatoriamente; no entanto, o sucesso da utilização do sistema é comprometido pela dificuldade de entendimento dos conceitos relacionados a mineração de regras de associação por parte dos usuários. Assim, na Seção 4.4 apresentamos o resultado de uma análise que fizemos dos desafios de entendimento relacionados à mineração de regras de associação que devem ser levados em conta pelo projetista no desenvolvimento da interface. Antes porém vamos mostrar alguns dados obtidos da análise dos *logs* de utilização do sistema.

Questão	Média	Desvio Padrão
Interface do Sistema	6,33	2,64
Utilização do Sistema	7,63	1,26
Quantidade de informação na tela	7,33	1,91
Arranjo das informações nas telas	6,93	1,62
Terminologia utilizada no sistema	6,80	2,43
As etapas para executar uma tarefa seguem uma seqüência lógica	7,75	1,34
As informações necessárias estão disponíveis para a realização da tarefa	6,44	2,19

Tabela 4.2: Médias das notas das questões fechadas do questionário.

4.3 Caracterização

Nesta seção, apresentamos alguns resultados de caracterização da utilização do Tamanduá com base nos *logs* do sistema, gerados tanto durante os treinamentos quanto fora deles. Esses resultados serão úteis na próxima seção, para sustentar alguns dos pontos que levantamos durante a observação do comportamento dos usuários nos cursos. Os *logs* foram gerados no período que vai de 05 de março a 16 de maio de 2005. Durante esse período, o sistema foi usado por 140 usuários distintos, em 793 sessões. Foram criadas 1712 tarefas no mesmo período (Tabela 4.3).

Número de Usuários	140
Número de Sessões	793
Número de Tarefas	1712

Tabela 4.3: Estatísticas extraídas dos *logs* no período entre 05 de março e 16 de maio de 2005.

Os gráficos a seguir mostram outras estatísticas da utilização do sistema: a Figura 4.4 mostra a distribuição dos usuários pelo número de sessões iniciadas por cada um deles; a Figura 4.5 mostra a distribuição das sessões de acordo com as suas durações; a Figura 4.6 mostra a distribuição das sessões de acordo com o número de tarefas executadas em cada uma delas.

Esses gráficos revelaram uma grande variabilidade no comportamento dos usuários do sistema. Em função disso, decidimos utilizar uma técnica de análise de agrupamentos para encontrar perfis representativos de sessões de usuários. Nesta análise, identificamos 3 perfis de sessões bem distintos: as sessões em que os usuários criam muitas tarefas, que representam cerca de 10% do total; as sessões em que os usuários criam poucas tarefas e visualizam superficialmente os resultados, que representam em

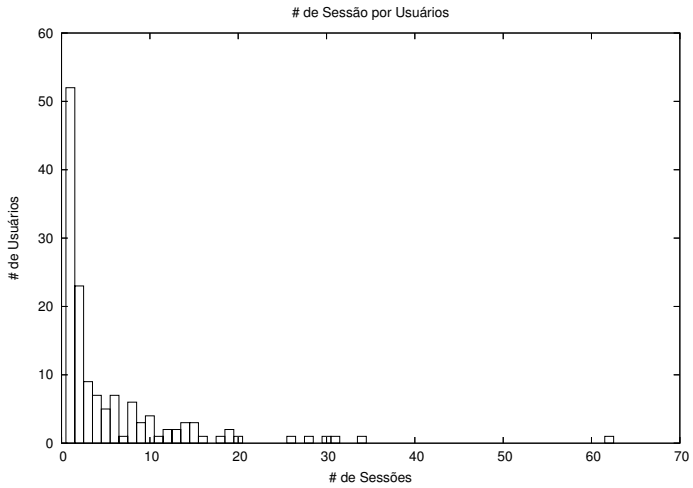


Figura 4.4: Distribuição dos usuários pelo número de sessões.

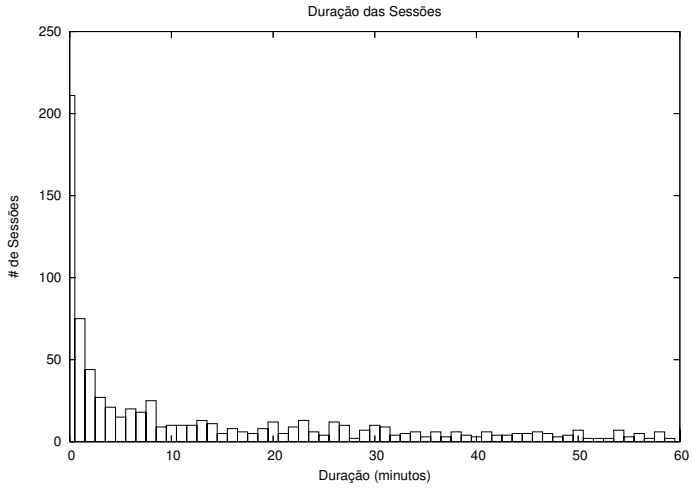


Figura 4.5: Distribuição das sessões de acordo com as suas durações.

torno de 25% do total; e as sessões em que os usuários realizam intensas visualizações dos resultados, representando cerca de 65% do total. Esses dados revelam importantes características do comportamento dos usuários de mineração de regras de associação. Os usuários criam inicialmente algumas tarefas, informando valores baixos para os parâmetros de suporte e confiança mínimos, o que leva à geração de quantidades muito grandes de regras. Em sessões subseqüentes, os usuários analisam os resultados, o que em função do volume de regras geradas consome muito tempo. Esta hipótese enfatiza a necessidade da criação de interfaces que facilitem a análise de grandes volumes de regras. Na próxima seção apresentamos outros resultados que reforçam esta hipótese.

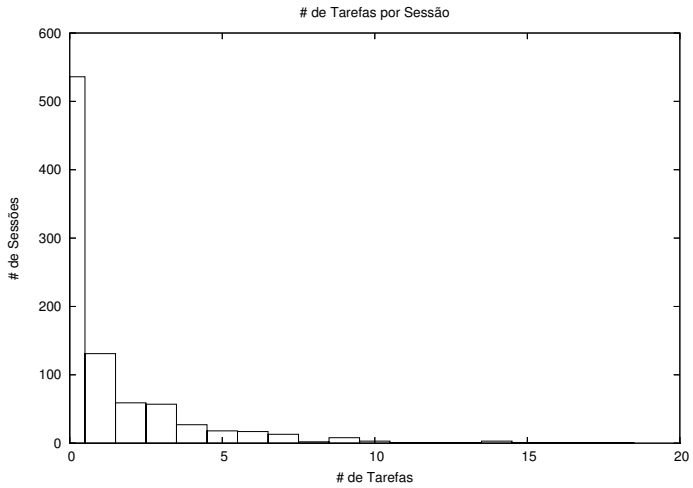


Figura 4.6: Distribuição das sessões de acordo com o número de tarefas criadas.

4.4 Desafio de Comunicabilidade

O nosso segundo objetivo nesta dissertação era auxiliar os usuários no entendimento das regras de associação e de todos os conceitos envolvidos. Esse objetivo se reforçou à medida em que realizamos as experiências com usuários: ficou cada vez mais claro que o maior problema em visualização de regras de associação estava relacionado ao entendimento por parte dos usuários dos conceitos específicos desta técnica. A seguir, apresentamos os principais aspectos relacionados aos sistemas de mineração de regras de associação que requerem conhecimento específico da área e que devem ser comunicados pelos projetistas aos usuários para possibilitar um melhor entendimento da solução proposta e assim possibilitar uma melhor utilização desses sistemas (Almir et al., 2005). Ao apresentar esses aspectos, explicamos em que eles consistem, quais as conseqüências de não comunicá-los eficientemente e o que torna particularmente difícil essa comunicação. Por uma questão de conveniência, organizamos esses aspectos em torno de três temas básicos: (1) visualização das regras, (2) conceitos relacionados às regras e (3) processo de geração das regras.

4.4.1 Visualização das Regras

O primeiro tema que vamos analisar compreende os aspectos relacionados à estratégia utilizada para exibição do conjunto de regras e aos mecanismos de foco oferecidos ao usuário. Estes aspectos são os que estão mais relacionados com a grande maioria dos trabalhos apresentados na Capítulo 2.

4.4.1.1 Aspecto 1: Paradigma Visual

O paradigma visual corresponde à estratégia utilizada pelo projetista para exibição do conjunto de regras. Como vimos no Capítulo 2, existem três grandes abordagens para exibição das regras: na primeira abordagem, o conjunto de regras é organizado em uma matriz antecedente×conseqüente; na segunda, o conjunto de regras é organizado na forma de um grafo, com as regras representadas nas arestas; e na terceira ele é organizado em uma matriz suporte×confiança. Além disso, cada uma dessas abordagens pode utilizar ainda o conceito de *itemset lattice*, organizando as regras de forma hierárquica. Esses paradigmas estão diretamente relacionados com os conceitos de mineração de regras de associação: antecedente, conseqüente, suporte, confiança, demais métricas de interesse, hierarquia e relacionamentos. Desta forma, eles requerem que o usuário entenda esses conceitos para que possam ser utilizados com eficiência. No sistema Tamanduá, tivemos indicadores das dificuldades enfrentadas por usuários sem formação em TI e os conseqüentes impactos sobre a utilização que fizeram.

A primeira versão do Tamanduá utilizava uma matriz antecedente×conseqüente para organizar as regras, além de utilizar o conceito de *itemset lattice*. Por expor esses conceitos específicos de regras de associação sem comunicar o seu significado de forma adequada, esta versão provocou uma forte resistência por parte dos primeiros usuários do sistema, que tinham que absorver uma quantidade muito grande de conhecimento antes de poder sequer começar a interagir. Esta dificuldade motivou o desenvolvimento da segunda versão do sistema. A versão atual do Tamanduá utiliza uma matriz suporte×confiança para organizar o conjunto de regras geradas, representando cada regra por um pequeno quadrado azul. Embora esta versão ainda exija o conhecimento de conceitos específicos, como suporte e confiança (além do próprio conceito de regra de associação), ela oferece uma barreira de entrada menor: transmitindo ao usuário o que representa cada quadrado e quais as regiões onde as regras potencialmente mais interessantes se encontram ele é capaz de começar a interagir. À medida em que ele começa a inspecionar as regras, novos conceitos vão sendo introduzidos, mas nesse ponto ele já venceu a resistência inicial.

Assim, na decisão sobre que paradigma visual a ser utilizado, o projetista deve considerar os conceitos técnicos que cada paradigma requer que os usuários conheçam para que eles possam interagir com o sistema. Assim, ele pode definir como a interface (no papel de seu preposto) vai apresentar e explicar aos usuários estes conceitos. O projetista pode até decidir que estes conhecimentos devem ser adquiridos previamente pelos usuários, e neste caso a interface deveria transmitir esta decisão aos usuários, deixando claro que conceitos ele deve conhecer de antemão e se for o caso comunicar até mesmo onde se espera que ele seja adquirido (e.g., tutoriais sobre o sistema, manuais

do sistema, material didático sobre mineração de dados etc.).

4.4.1.2 Aspecto 2: Seleção de Sub-conjuntos de Regras

Além de permitir ao usuário visualizar as regras geradas e suas características, o sistema de mineração de regras de associação deve também permitir ao usuário selecionar um sub-conjunto de regras que seja mais interessante para ele. Assim, o projetista deve também comunicar ao usuário através da interface que especificações sobre o conjunto de regras geradas ele pode fazer e como fazê-lo. O problema aqui é semelhante ao problema enfrentado pelos projetistas de interfaces de consulta em bancos de dados ou em sistemas de recuperação da informação. Quanto maior a flexibilidade oferecida pela interface de consulta, maior a dificuldade do usuário em utilizar a interface. Em um extremo, podemos permitir ao usuário a definição de expressões lógicas complexas, inclusive com parênteses para indicar precedência. Neste caso, a dificuldade de uso poderá ser grande, mas a flexibilidade será maximizada. No outro extremo, podemos limitar as opções do usuário, oferecendo controles tais como *sliders* e *check-boxes* para seleção dos itens, e uma configuração fixa das combinações lógicas entre os controles. Neste caso, a facilidade de uso será maior, mas dificultará ao usuário fazer consultas mais complexas.

Os sistemas de mineração de regras de associação adicionam a esse problema a questão da estrutura das regras de associação. Assim, além de definir quais itens o interessam, o usuário poderia definir em que lado da regra ele quer que um determinado item esteja presente, ou qual o número de itens uma regra deve ter no antecedente ou no conseqüente para ser considerada interessante. Para isso, é importante que o usuário entenda não apenas a estrutura da regra (e.g., que a regra é formada por um antecedente e um conseqüente), mas também o que significa um item estar presente de um lado ou do outro.

No sistema Tamanduá a seleção de sub-conjuntos de regras é feita através de um mecanismo de filtro que permite que o usuário especifique os itens que ele considera interessantes. O usuário pode definir também de que lado da regra ele quer que esses itens ocorram. Durante as nossas experiências, pudemos perceber que esse mecanismo de filtro não estava bem comunicado ao usuário. Em geral, um item deve ser especificado no antecedente quando se deseja analisar as suas conseqüências e no conseqüente quando se deseja analisar as suas causas. A maioria dos usuários se mostrou confusa em tomar essa decisão, seja por não entender essa diferença, seja por não estar interessado em analisar relações de causa e conseqüência. Além disso, dependendo do lado que o usuário escolher, a quantidade de regras exibidas pode ser drasticamente diferente. Por exemplo, isto pode acontecer pela restrição de confiança mínima especificada, já que

um mesmo conjunto de itens pode gerar regras com confianças diferentes. Considere por exemplo as seguintes regras:

[Pão], [Manteiga] ⇒ [Leite] (80.00, 50.00)
[Pão] ⇒ [Leite], [Manteiga] (40.00, 50.00)

Ambas, como era de se esperar, apresentam o mesmo valor de suporte (50%), já que elas são geradas a partir do mesmo conjunto de itens (Pão, Manteiga e Leite). No entanto, como a confiança dessas regras é diferente, se o usuário especificar uma confiança mínima de 60%, apenas a primeira regra será gerada (já que a segunda ficaria abaixo deste valor). Assim, se ele filtrar as regras pela presença do item Manteiga no antecedente, pelo menos a primeira regra será exibida. Por outro lado, se ele filtrar as regras pela presença do item Manteiga no conseqüente, pode ser que nenhuma regra seja exibida. Se o usuário não entender as razões para esta diferença, ele tentará criar hipóteses para explicar este comportamento. Uma hipótese poderia ser de que a presença de itens no antecedente sempre encontra mais regras. Se ele se basear nesta hipótese ele poderá sempre definir esta opção para selecionar as regras e talvez nunca perceber que sua hipótese era falsa. O custo disto para o usuário é que em algumas situações ele poderia não tomar conhecimento de regras que poderiam ser interessantes para ele, presumindo que elas não existiam (e não que ele não as encontrara).

4.4.2 Conceitos Relacionados às Regras

O segundo tema realça a importância do entendimento das regras de associação e métricas de interesse pelos usuários. Como vimos no Capítulo 2, existem muito poucos trabalhos relacionados a este tema além do trabalho de Hofmann et al. (2000).

4.4.2.1 Aspecto 3: Conceito de Regra de Associação

O conceito de regra de associação não é um conceito trivial. Cada regra de associação representa uma possível correlação entre itens de uma base de dados. Possível porque o fato de existir uma regra de associação entre dois ou mais itens não significa necessariamente que eles estejam correlacionados, como veremos mais adiante. Vamos considerar a regra que usamos de exemplo anteriormente:

[Pão], [Manteiga] ⇒ [Leite] (80.00, 50.00)

Esta regra indica uma possível correlação entre a compra de Pão, Manteiga e Leite. Como vimos, ela nos diz que os itens Pão, Manteiga e Leite são comprados juntos com uma freqüência de 50%, e que 80% das compras que incluem Pão e Manteiga

também incluem Leite. Esta última porcentagem corresponde também à chance de acerto de uma previsão da compra de Leite dado que já ocorreu a compra de Pão e Manteiga. É muito importante que o usuário compreenda essas informações, porque além delas serem interessantes por si só, elas são fundamentais para a compreensão dos demais conceitos utilizados no sistema. Além disso, se o usuário não interpretá-las corretamente, corre-se o risco de ele utilizá-las de forma equivocada.

Na literatura, convencionou-se utilizar alguns termos que nem sempre são os mais adequados do ponto de vista conceitual, e que embora façam sentido do ponto de vista de implementação, nem sempre facilitam o entendimento do que significam. Acontece que esses termos acabam sendo utilizados nas interfaces, sem um apoio maior ao seu entendimento e, logo, dificultando a compreensão das mesmas por usuários leigos. Um exemplo típico é o termo *item*, conceito fundamental em regras de associação. Um item corresponde a um par atributo-valor, como *Sexo=Feminino*, *Idade=27* ou *Pão=Verdadeiro* (que pode ser abreviado simplesmente como *Pão*). Os itens formam os conjuntos de itens, que por sua vez formam o antecedente e o conseqüente, de cuja relação se forma uma regra. O suporte é calculado pela freqüência do conjunto de itens formado da união do antecedente e do conseqüente, e a confiança é dada pela freqüência do conjunto de itens que formam o conseqüente, no espaço definido pelo conjunto de itens que forma o antecedente. Uma sugestão para o projetista seria buscar termos para representar estes conceitos na interface que poderiam ajudar o usuário a entendê-los.

Assim, diversos conceitos são comunicados de forma inadequada, levando o usuário a conclusões enganosas. Por exemplo, uma regra com suporte e confiança altos muitas vezes leva o usuário a achar que aquela regra revela uma correlação entre os itens, quando isso nem sempre é verdade. O suporte geralmente é usado apenas como parâmetro de poda, para evitar que muitas regras sejam geradas. A confiança por sua vez não deve ser usada como medida de correlação por si só, já que para isso existem as outras métricas de interesse. Outro engano comum é achar que uma regra sempre deve expressar uma relação de causa e conseqüência. Considere por exemplo a seguinte regra hipotética, que poderia ter sido minerada em uma base de dados de ocorrências policiais de uma metrópole qualquer:

$$[\text{Ocorrência=Homicídio}] \Rightarrow [\text{Dia=Domingo}]\ (23.00,\ 0.60)$$

Esta regra traz algumas informações interessantes. Ela mostra que 23% dos homicídios cometidos nesta metrópole ocorrem no domingo e que 0,6% das ocorrências registradas nesta metrópole são homicídios cometidos no domingo. Mas o usuário deve entender que essas informações não são suficientes para concluir que existe uma correlação entre a ocorrência de homicídios e o dia da semana. E também não faria o menor sentido concluir que determinada quantidade de ocorrências de homicídios seja

a causa para o dia da semana ser domingo. Se os conceitos de regras de associação não forem comunicados adequadamente, o usuário pode ser levado a chegar a conclusões desse tipo.

No Tamanduá, os usuários, mesmo depois de terem feito o curso de introdução aos conceitos, são freqüentemente encontrados tentando entender o significado básico de uma regra, geralmente sem sucesso. A maior dificuldade dos usuários encontra-se em entender o conceito de confiança. Isto ocorre talvez pelo fato do conceito de confiança ser explicado a partir dos conceitos de antecedente e conseqüente, que normalmente já são difíceis de serem absorvidos pelo usuário. Quando o usuário não consegue entender bem esses conceitos, ele os utiliza de forma inadequada. Pudemos observar que os usuários do Tamanduá freqüentemente invertem os conceitos de suporte e confiança, ou interpretam o suporte como a freqüência do antecedente na base, quando sabemos que o suporte corresponde à freqüência da união do antecedente com o conseqüente. O resultado disso é que o usuário pode acabar interpretando erroneamente as informações sendo apresentadas pelo sistema. Além disso, ele terá também dificuldade em compreender outros conceitos do sistema que dependam da compreensão desses conceitos básicos.

4.4.2.2 Aspecto 4: Métricas de Interesse

Dissemos anteriormente que uma regra com suporte e confiança altos não necessariamente indica uma correlação entre os itens. Para medir essa correlação, precisamos de outras métricas de interesse. Na literatura, são encontradas dezenas dessas métricas, algumas mais adequadas a determinadas situações que as outras. Uma métrica de interesse bastante popular é o *lift*, que dá uma medida do quanto a confiança de uma regra é surpreendente com relação ao que era esperado. No nosso exemplo, a confiança de 80% indica que 80% das compras que incluíram Pão e Manteiga também incluíram Leite. Embora essa confiança pareça alta, não podemos afirmar isso com certeza sem olharmos a freqüência da compra de Leite na base de dados. Se 80% de todas as compras efetuadas na padaria incluíram Leite, então a confiança de 80% já era esperada, e a regra não teria trazido nenhuma informação surpreendente. Por outro lado, se apenas 40% de todas as compras efetuadas na padaria incluíram Leite, então a confiança de 80% é o dobro da esperada, indicando que a compra de Pão e Manteiga influencia positivamente na compra de Leite, o que é uma informação surpreendente. O *lift* é dado pela razão entre a confiança da regra e a confiança que seria esperada. Se a confiança esperada era de 80% e a confiança da regra foi de 80%, o *lift* é 1. Da mesma forma, se a confiança esperada era de 40% e a confiança da regra foi de 80%, o *lift* é 2. Quanto mais o *lift* divergir do valor 1, maior será a intensidade da correlação expressa

pela regra e mais surpreendente ela será. Valores de *lift* menores que 1 indicam uma correlação negativa e valores de *lift* maiores que 1 indicam uma correlação positiva.

Algumas métricas são mais intuitivas do que as outras. O *lift* é um exemplo de uma métrica intuitiva: basta que o usuário entenda o conceito de confiança para que possa ser capaz de compreendê-la. Outras métricas, como o *leverage* ou a convicção são mais complicadas, embora possam ser explicadas de forma análoga ao *lift*. Existem ainda métricas que são muito mais difíceis de serem compreendidas. No Tamanduá, as regras são organizadas em uma matriz suporte×confiança, mas o usuário tem a opção de utilizar os eixos para representar 3 outras métricas de interesse em substituição ao suporte e à confiança: *lift*, *leverage* e convicção. Nas nossas experiências com os usuários do Tamanduá, pudemos observar que o *lift* era de fato a métrica mais facilmente compreendida dessas três, o que talvez justifique o fato dela ser a mais usada pelos usuários. A Tabela 4.4 mostra a freqüência com que os usuários do Tamanduá utilizam cada métrica em um dos eixos. O par suporte×confiança é o mais popular, o que se justifica tanto pelo fato de ser esta a configuração *default* quanto pelo fato dessas serem as métricas de mais fácil compreensão. Mas os pares suporte×*lift* e confiança×*lift* são bastante populares também, o que não ocorre com os pares envolvendo *leverage* e convicção.

Métrica	Freqüência
Confiança	85,01%
Suporte	66,03%
Lift	40,70%
Leverage	3,30%
Convicção	2,66%

Tabela 4.4: Freqüência de utilização das métricas de interesse no eixos pelos usuários do Tamanduá.

O usuário consegue usar o sistema apenas com as métricas de suporte e confiança e pode chegar a regras interessantes que lhes são úteis. No entanto, o impacto de ele não entender as outras métricas e nem quando usá-las é que ele acaba não utilizando recursos que estão disponíveis e que em muitas situações poderiam resultar em um conjunto de regras mais interessante ou significativo.

4.4.3 Processo de Geração de Regras

O terceiro tema que vamos abordar está relacionado ao processo de geração das regras de associação. Este tema envolve as decisões que devem ser tomadas pelos usuários antes que o sistema inicie a execução dos algoritmos que irão gerar as regras. Essas

decisões influenciam tanto no tempo que o algoritmo demora para executar quanto nos resultados que ele gera. Assim, é importante comunicar aos usuários que decisões são estas e seu impacto sobre o uso do sistema.

4.4.3.1 Aspecto 5: Definição dos Parâmetros

Os algoritmos de mineração de regras de associação geralmente exigem que o usuário defina alguns parâmetros iniciais para que eles possam ser executados. Os dois parâmetros mais tradicionais desses algoritmos são o suporte e a confiança mínimos. O usuário deve fornecer o valor mínimo de suporte que uma regra deve apresentar para que ela seja gerada, o mesmo valendo para a confiança. Os valores mais adequados para esses parâmetros dependem da base de dados que vai ser minerada e de algumas premissas do usuário. Embora mesmo sendo difícil para um especialista em mineração de regras de associação encontrar o valor ideal para esses parâmetros, é possível se determinar uma faixa de valores adequada. Se o usuário informar valores fora dessa faixa, o sistema pode gerar um número muito grande de regras, dificultando a análise, ou gerar um número de regras insuficiente, excluindo as regras que realmente interessam ao usuário. Assim, o ideal seria que a interface conseguisse transmitir ao usuário não apenas os conceitos, mas como determinar esta faixa de valores.

Quando o usuário não entende quais os valores que ele deve usar como parâmetros, ele utiliza valores aleatórios, o que implica em se ter sempre um custo associado à tentativa de encontrar valores adequados. Nas experiências com o Tamanduá, a grande maioria dos usuários não sabia como decidir que valores usar como parâmetros e muitos deles se sentiam frustrados por isso. A análise dos *logs* do Tamanduá revelou que os usuários utilizam geralmente valores bem parecidos de suporte e confiança mínimos, o que demonstra que eles também não entendem bem a diferença entre os dois parâmetros. Os *logs* revelaram também que um comportamento bem típico dos usuários é utilizar valores de suporte e confiança mínimos inicialmente altos, reduzindo gradativamente esses valores até culminar com uma redução brusca. Em outras palavras, os usuários adotavam uma estratégia de tentativa e erro até obter um número de regras que eles consideravam satisfatório.

Além da frustração que isto pode causar ao usuário, este não entendimento pode acarretar em outros custos para ele. Por exemplo, uma vez encontrados estes valores satisfatórios para em um determinado contexto, observa-se que eles tendem a repetilos com outras bases de dados, ou com visões diferentes da mesma base. No entanto, os valores ideais para uma determinada situação, normalmente não se aplicam a outras. Assim, este comportamento indica que além de não entenderem como chegar a uma faixa de valores adequada, eles não compreendem a dependência destes valores

do contexto em que são utilizados. Desta forma, eles não têm opção se não passar novamente pelo processo de tentativa e erro novamente para conseguir determinar os valores adequados a cada nova situação.

4.4.3.2 Aspecto 6: Escolha dos Atributos

Um outro aspecto relacionado à geração das regras que merece ser mencionado tem a ver com a escolha dos atributos a serem minerados na base de dados. Muitas bases de dados possuem alguns atributos que são redundantes ou parcialmente redundantes. Por exemplo, numa base de compras os atributos "código do produto" e "nome do produto" em geral são redundantes, já que cada código corresponde a um único produto (e.g., o código "123" corresponde ao produto "Mouse XYZ") . Já os atributos "nome do produto" e "categoria do produto" são parcialmente redundantes, já que cada produto é de uma única categoria (e.g., o produto "Mouse XYZ" pertence à categoria "Periféricos"). Quando o usuário seleciona atributos redundantes ou parcialmente redundantes, o sistema pode gerar regras óbvias, como as seguintes:

$$[\text{Código=123}] \Rightarrow [\text{Nome=Mouse XYZ}]\ (100.00,\ 1.00)$$
$$[\text{Nome=Mouse XYZ}] \Rightarrow [\text{Categoria=Periféricos}]\ (100.00,\ 1.00)$$

É óbvio que 100% dos produtos de código "123" são "Mouse XYZ", assim como é óbvio que 100% dos "Mouse XYZ" sejam "Periféricos". Como o sistema não tem como saber que os atributos são redundantes, essas regras irão aparecer em destaque, já que possuem uma confiança alta e um *lift* também alto. O *lift* da primeira regra por exemplo tem valor 100, indicando que a confiança da regra é 100 vezes maior que a freqüência do conseqüente (que nesse caso é 1%). Ou seja, o fato de sabermos que o código do produto em uma determinada compra é igual a "123" aumenta em 100 vezes a chance do nome do produto na mesma compra ser "Mouse XYZ", o que é óbvio.

O usuário normalmente entende os atributos e seus valores, uma vez que eles pertencem ao seu domínio. Porém se o usuário não entende como a seleção de atributos impacta a geração de regras, ele pode selecionar atributos redundantes, e obter regras óbvias. Conhecendo o domínio da base de dados, o usuário percebe que estas regras são óbvias, porém ele não entende por que o sistema está lhe dizendo que estas regras são potencialmente muito interessantes, quando de fato elas não são nada interessantes. Isto pode levar o usuário a perder a confiança no sistema e na sua capacidade de gerar informações que sejam interessantes para ele, e acabar optando por não utilizar o sistema. Nas experiências no Tamanduá, a análise dos *logs* indicou que 17,82% de todas as minerações efetuadas no sistema incluíram atributos redundantes. Alguns usuários nesta situação demonstraram irritação ao inspecionarem regras como essas.

4.5 Sumário

Apresentamos assim os resultados da avaliação das nossas estratégias de visualização de regras de associação. A estratégia de Visualização Estrutural se mostrou inferior à estratégia de Visualização por Métricas de Interesse. Esta última foi avaliada positivamente pelos usuários, que a consideraram fácil de usar. No entanto, a despeito da facilidade de uso associada ao paradigma visual utilizado, os usuários de mineração de dados enfrentam um outro problema que compromete a utilização do sistema: os conceitos associados à técnica precisam ser bem comunicados para que eles possam compreendê-los e utilizar o sistema satisfatoriamente.

Assim, levantamos os aspectos para os quais o projetista de um sistema de mineração de regras de associação deve ficar atento, uma vez que esses aspectos devem ser comunicados aos usuários. Usamos a teoria de engenharia semiótica para apoiar este estudo, já que as abordagens tradicionais de IHC não se mostraram adequadas. A engenharia semiótica entende a interface de um sistema como uma mensagem enviada pelo projetista ao usuário. Identificamos 6 aspectos que devem ser comunicados ao usuário e o custo para ele se esses aspectos não forem adequadamente comunicados. Vimos que as principais conseqüências são a dificuldade de encontrar as regras mais interessantes, a dificuldade de compreensão de uma regra e, em última instância, a perda da confiança no sistema.

Capítulo 5

Conclusão

A mineração de dados é uma nova área de pesquisa, que experimentou um desenvolvimento muito grande nos últimos anos, fruto de uma demanda por novas técnicas de análise capazes de lidar com grandes volumes de dados. Uma das mais conhecidas técnicas de mineração de dados, a mineração de regras de associação encontrou um vasto número de aplicações nos mais diferentes domínios, desde marketing até ciências sociais. No entanto, os sistemas de mineração de dados ainda requerem a presença de um especialista, seja qual for o domínio em que sejam utilizados.

Para reduzir a dependência desse especialista e permitir que tais sistemas sejam utilizados de maneira mais ampla e eficiente por usuários leigos, é necessário que as pesquisas em mineração de dados busquem um apoio maior na área de IHC. Existem basicamente dois problemas relacionados à interação humana em sistemas de mineração de regras de associação: em primeiro lugar, a mineração de regras de associação gera uma quantidade muito grande de resultados, o que torna necessária a criação de mecanismos que auxiliem o usuário a lidar com esses resultados para encontrar neles aquilo que realmente o interessa; em segundo lugar, a mineração de regras de associação exige do usuário um conhecimento dos conceitos específicos desta técnica para que ele possa tirar proveito dos resultados. O nosso objetivo nesta dissertação era criar uma interface para suportar o usuário, especialista ou leigo, na análise dos resultados de uma tarefa de mineração de regras de associação, o que significava abordar esses dois problemas de interação. Neste capítulo, apresentamos as nossas principais contribuições com relação a esses dois problemas e apontamos direções para possíveis trabalhos futuros.

Apresentamos duas estratégias para visualização de regras de associação. Estas estratégias representam propostas de solução para o problema de identificação das regras mais interessantes no conjunto de regras geradas. Para desenvolvê-las, fizemos uma análise dos usuários do sistema e das tarefas realizadas por eles (as quais deveriam

ser suportadas pela interface) e projetamos uma primeira versão, que em seguida passou por um ciclo de prototipagens e avaliações até chegar à versão atual. Esta versão possui as seguintes características principais:

- Utiliza o paradigma de matriz suporte×confiança, onde os eixos são usados para representar os valores de suporte e confiança das regras, e cada regra é representada no ponto dado pelo seu suporte e confiança;
- Utiliza a abordagem "visão geral primeiro, filtro e então detalhes sob demanda";
- Oferece um mecanismo de filtro que permite encontrar as regras pela presença de itens específicos no antecedente ou no conseqüente;
- Permite que as medidas representadas nos eixos (suporte e confiança) sejam substituídas por até três outras medidas: *lift*, *leverage* e convicção;
- Permite a visualização dos dados atômicos que são sumarizados por uma regra.

A nossa principal contribuição neste aspecto foi o desenvolvimento de duas estratégias bem distintas e a avaliação empírica dessas estratégias. Mostramos que a abordagem baseada em uma matriz suporte×confiança, que chamamos de Visualização por Métricas de Interesse, parece ser superior à abordagem baseada em uma matriz antecedente×conseqüente, que chamamos de Visualização Estrutural.

Como trabalhos futuros nesta linha, apostamos na incorporação dos novos avanços da área de algoritmos, como sumarização de regras e eliminação de redundância, ao paradigma visual. A incorporação desses conceitos irá exigir a utilização de alguma técnica de visualização de hierarquias. Um outro problema a ser tratado está relacionado à sobreposição de regras. Uma possibilidade é a utilização de *zoom*.

As interfaces que desenvolvemos serviram também para subsidiar um estudo de caracterização dos desafios de entendimento relacionados a mineração de regras de associação, que constituem o segundo problema mencionado anteriormente. Com o objetivo de apoiar projetistas de sistemas de mineração de regras de associação no entendimento destes desafios, nesta dissertação identificamos seis aspectos técnicos relativos às regras de associação que devem estar bem comunicados aos usuários para que eles consigam entender e utilizar estes sistemas. Para cada aspecto explicamos os conceitos envolvidos, a sua importância para o uso do sistema, e os potenciais impactos quando não estão bem comunicados para o usuário. Estes aspectos foram definidos a partir da literatura de interação em sistemas de mineração de dados e da experiência dos usuários observada no Projeto Tamanduá.

A identificação destes aspectos contribui para projetistas de sistemas de mineração de regras de associação, uma vez que chama a sua atenção para o conhecimento técnico

necessário para o usuário e sua importância para o sucesso do sistema. Assim, ele contribui para que o projetista reflita sobre as questões levantadas e tome decisões informadas de que conhecimento técnico ele deve transmitir ao usuário e como fazê-lo através da interface. Para a engenharia semiótica os aspectos identificados nesta dissertação e as explicações associadas a ele constituem uma ferramenta epistêmica para o projetista, uma vez que eles contribuem para o entendimento do projetista sobre aspectos do problema sendo resolvido e o ajudam a refletir sobre o seu conhecimento.

Os aspectos levantados também podem contribuir para a avaliação de interfaces de sistemas de mineração de regras de associação. O avaliador da interface pode utilizar estes aspectos para inspecionar a qualidade da comunicação destes conhecimentos técnicos através da interface. Eles podem ser úteis também para o avaliador projetar tarefas para um teste com usuários ou formular questões para uma entrevista, que lhe permitam ter indicadores sobre o quanto os usuários adquiriram ou não os conhecimentos necessários, representações na interface que facilitaram ou dificultaram o seu entendimento e como afetaram o seu desempenho. Estes aspectos podem auxiliar ainda na análise de dados coletados a partir do uso do sistemas, oferecendo insumos para que os avaliadores entendam comportamentos relacionados com os pontos discutidos nesta dissertação.

Como vimos no Capítulo 2, as pesquisas sobre interação para mineração de regras de associação têm focado principalmente os paradigmas de visualização necessários. No entanto, Thomas et al. (1999) chamam a atenção para a necessidade de novos paradigmas de interação para lidar com a enorme quantidade de dados que se tem disponível hoje. Os aspectos levantados nesta dissertação contribuem neste esforço, uma vez que apontam para aspectos de qualidade da interação que deveriam estar presentes nestas novas propostas.

A teoria da engenharia semiótica argumenta que se o usuário entende a visão de *design* do projetista de um sistema, ele tem melhores chances de usar com eficiência o sistema. Os sistemas de mineração de regras de associação requerem que os usuários adquiram um conhecimento técnico sobre a solução para que consigam utilizá-la. Assim, este tipo de sistema e o desafio de comunicabilidade apresentado nesta dissertação fortalecem o ponto defendido pela engenharia semiótica, uma vez que neste caso a comunicabilidade não apenas amplia a capacidade do usuário fazer um uso eficiente do sistema, mas é fundamental para que ele consiga utilizá-lo.

Os resultados deste trabalho trazem contribuições práticas para o projeto Tamanduá, uma vez que os aspectos levantados identificaram a necessidade da melhoria da comunicabilidade da interface sobre alguns conhecimentos técnicos. Um primeiro passo nesta direção será definir como alterar a interface para melhorar sua comunicabilidade, nos pontos que já identificamos que os usuários apresentam dificuldade. A partir de

uma inspeção utilizando os aspectos identificados, verificaremos se existem pontos críticos onde seria desejável uma avaliação formal utilizando o método de avaliação de comunicabilidade (Prates et al., 2000). Finalmente, para melhorar a comunicabilidade de alguns dos conceitos técnicos, nos parece necessário fazer um melhor uso do sistema de ajuda do Tamanduá. Assim, usando como base o modelo de sistema de ajuda proposto com base na engenharia semiótica (Silveira et al., 2003) e levando em consideração que o sistema deverá assumir um papel mais tutorial em relação aos conceitos específicos de mineração de dados, será feita uma nova versão para o sistema de ajuda do Tamanduá.

Esta dissertação levanta novas questões para investigação. Aqui discutimos o desafio de comunicabilidade para o projetista transmitir a usuários leigos conceitos técnicos da área de mineração de dados necessários para interagir com o sistema. Mostramos a necessidade de os usuários entenderem estes conceitos para conseguirem utilizar o sistema. No entanto, estes usuários leigos muitas vezes têm diferentes perfis e usam o sistema com diferentes objetivos. Assim, um ponto a ser investigado é de se e como estes perfis ou objetivos influenciam a profundidade deste conhecimento técnico necessário pelo usuário. Por exemplo, podemos pensar que um sociólogo utilizando o sistema para levantamento de dados que apóiem uma teoria sociológica precisa entender muito bem os conceitos apresentados para poder conseguir encontrar os dados, ou chegar à conclusão que eles não existem (e não que ele não foi capaz de encontrá-los). Por outro lado, um auditor que utilize o sistema para identificar candidatos interessantes para uma auditoria talvez possa ter um conhecimento mais superficial dos conceitos técnicos.

Finalmente, este trabalho focou em mineração de regras de associação, mas conforme apresentamos no Capítulo 1, existem outras técnicas de mineração. Desta forma seria interessante investigar se e quais dos aspectos levantados neste trabalho se aplicam a outras técnicas, como classificação e análise de agrupamentos. Também seria interessante ver se estas novas técnicas envolvem outros aspectos não relevantes para regras de associação.

Apêndice A

Questionário para Levantamento do Perfil dos Usuários

Este questionário faz parte de uma pesquisa que estamos realizando para analisar o perfil dos usuários potenciais do sistema Tamanduá. A sua colaboração é fundamental para que a pesquisa sirva para melhorar a usabilidade do sistema. Preste atenção às orientações colocadas entre colchetes ao longo do questionário. Qualquer dúvida, esclarecimento ou reclamação, entre em contato com a equipe do Projeto Tamanduá, através do e-mail: suporte@tamandua.speed.dcc.ufmg.br.

Vamos começar com algumas perguntas sobre sua idade, sexo, escolaridade e profissão.

P1) Em que ano você nasceu?
Ano [Anotar]: _____

P2) Qual o seu sexo?
(1) Feminino
(2) Masculino

P3) Qual o seu grau de instrução?
(1) 1º grau incompleto
(2) 1º grau completo
(3) 2º grau incompleto
(4) 2º grau completo
(5) Superior incompleto
(6) Superior completo
(7) Pós-graduação incompleta

(8) Pós-graduação completa

P4) Qual é o curso que você está fazendo atualmente ou o último curso que você completou?

Curso [Anotar]: _____

P5) Qual é a sua profissão?

Profissão [Anotar]: _____

P6) Há quanto tempo você se encontra nessa profissão?

(1) Menos de 1 ano
(2) Entre 1 e 2 anos
(3) Entre 2 e 4 anos
(4) Mais de 4 anos

P7) Você utiliza óculos ou lentes de contato?

(1) Não, pois não preciso
(2) Não, mas deveria
(3) De vez em quando, quando me lembro
(4) Sim, quando leio
(5) Sempre, senão não enxergo nada

As perguntas a seguir são sobre utilização do computador.

C1) Há quanto tempo você utiliza o computador?

(1) Menos de 1 ano
(2) Entre 1 e 2 anos
(3) Entre 2 e 4 anos
(4) Mais de 4 anos

C2) Em que local(is) você costuma utilizar o computador?

(1) Em casa
(2) No trabalho
(3) Na escola
(4) Outro(s). Qual(is)? [Anotar]: _____

C3) Em média, quantas horas por dia você utiliza o computador?

(1) Menos de 1 hora

(2) Entre 1 e 3 horas

(3) Entre 3 e 6 horas

(4) Mais de 6 horas

C4) Qual(is) programas você costuma utilizar fora do trabalho?

(1) E-mail

(2) Internet Banking

(3) Mensagens instantâneas (ICQ, MSN Messenger, Yahoo! Messenger, etc.)

(4) Chat, Blogs, Orkut

(5) Outro(s). Qual(is)? [Anotar]: _____

C5) Como você prefere aprender a utilizar um programa de computador?

(1) Lendo, gosto de rever toda a documentação antes de começar a usar um programa

(2) Observando e fazendo perguntas àqueles que já sabem utilizar

(3) Explorando por conta própria, pois aprendo melhor com meus próprios erros e descobertas

(4) Através de cursos e treinamento formais com instrutores

(5) Outro(s). Quais(s)? [Anotar]: _____

Agora algumas perguntas sobre a sua experiência com estatística e mineração de dados.

MD1) Você utiliza ou já utilizou alguma técnica de estatística?

(1) Sim

(2) Não [Vá para MD3]

MD2) Qual(is) técnica(s) de estatística você utiliza ou já utilizou?

(1) Freqüências, Gráficos de Barra, Média, Variância, etc.

(2) Probabilidade e Probabilidade Condicional

(3) Distribuições de Probabilidade (Binomial, Poisson, Normal, etc.)

(4) Análise de Correlação Linear

(5) Modelos de Regressão Linear

(6) Análise de Variância (ANOVA)

(7) Testes de Hipóteses (Qui-quadrado, t-Student, etc.)

(8) Análise de Séries Temporais

(9) Análise Multivariada

(10) Simulação

(11) Outra(s). Qual(is)? [Anotar]: _____

MD3) Você utiliza ou já utilizou algum programa estatístico?

(1) Sim

(2) Não [Vá para MD5]

MD4) Qual(is) programa(s) estatístico(s) você utiliza ou já utilizou?

(1) SAS

(2) SPSS

(3) S-PLUS

(4) R

(5) WinBUGS

(6) Stata

(7) EViews

(8) Minitab

(9) Outro(s). Qual(is)? [Anotar]: _____

MD5) Você utiliza ou já utilizou algum programa de mineração de dados (além do Tamanduá)?

(1) Sim

(2) Não [Vá para D1]

MD6) Qual(is) programa(s) de mineração de dados você utiliza ou já utilizou?

(1) Clementine

(2) DBMiner

(3) IBM Intelligent Miner

(4) Insightful Miner

(5) Oracle Data Mining

(6) Purple Insight MineSet

(7) SAS Enterprise Miner

(8) Weka

(9) Outro(s). Qual(is)? [Anotar]: _____

A seguir, algumas perguntas sobre a utilização do Tamanduá na sua organização.

D1) Qual a sua opinião com relação à utilidade do sistema Tamanduá para a sua organização?

(1) Acho que ele tem grande potencial, podendo ser útil em muitos problemas

(2) Acho que ele vai ser útil em apenas alguns problemas
(3) Acho que ele não vai ser útil
(4) Não sei responder
(5) Não se aplica

D2) Você acha que a sua opinião é importante na decisão de se utilizar o sistema Tamanduá na sua organização?
(1) Sim
(2) Não
(3) Não sei responder
(4) Não se aplica

D3) Você acha que a sua opinião será levada em conta nessa decisão?
(1) Sim
(2) Não
(3) Não sei responder
(4) Não se aplica

Obrigado pela sua colaboração!

Apêndice B

Lista de Tarefas Usada nas Avaliações com Usuários

Lista de Tarefas – Sistema Tamanduá

Abaixo temos 13 tarefas que devem ser executadas por você na ordem em que se encontram, utilizando o sistema.

Lembre-se:

- Verbalize suas dúvidas, pois isto nos ajudará a anotar a ocorrência e razão dos problemas.
- É o sistema que está sendo avaliado e não você.

Tarefa 1	Entrar no sistema Tamanduá utilizando o login e senha fornecidos
Tarefa 2	Criar uma nova Tarefa Fornecer o nome e descrição da Tarefa Escolher a base a ser utilizada: copom_PMMG Escolher os atributos que considerar interessantes Escolher o algoritmo Apriori Atribuir valores para as medidas de suporte e confiança
Tarefa 3	Executar a Tarefa
Tarefa 4	Visualizar os resultados da tarefa executada
Tarefa 5	Modificar o valor do suporte inicialmente atribuído
Tarefa 6	Visualizar novamente os resultados

Tarefa 7	Encontrar a regra com o maior suporte
Tarefa 8	Encontrar a regra com a maior confiança
Tarefa 9	Encontrar a regra com o maior lift
Tarefa 10	Encontrar a regra com a maior leverage
Tarefa 11	Selecionar 3 regras que julgar interessantes
Tarefa 12	Acessar o manual do sistema
Tarefa 13	Sair do sistema, respondendo ao questionário de satisfação

Referências Bibliográficas

Aggarwal, C. C. (1998). A human-computer cooperative system for effective high dimensional clustering. In *Proceedings of IEEE ICDE 1998 (14th International Conference on Data Engineering)*, pp. 221–226.

Aggarwal, C. C. e Yu, P. S. (1998). Online generation of association rules. In *ICDE '98: Proceedings of the Fourteenth International Conference on Data Engineering*, pp. 402–411, Washington, DC, USA. IEEE Computer Society.

Agrawal, R.; Imielinski, T. e Swami, A. (1993). Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–216. ACM Press.

Almir, F.; Tuler, E.; Rocha, L.; Prates, R. e Meira, W. (2005). Desafios de comunicabilidade para o projeto de sistemas de mineração de regras de associação. Submetido para o CLIHC 2005: Segunda Conferência Latino-Americana de Interação Humano-Computador.

Ankerst, M.; Ester, M. e Kriegel, H.-P. (2000). Towards an effective cooperation of the user and the computer for classification. In *Proceedings of ACM SIGKDD 2000 (6th International Conference on Knowledge Discovery and Data Mining)*, pp. 179–188.

Bayardo, R. J.; Agrawal, R. e Gunopulos, D. (2000). Constraint-based rule mining in large, dense databases. *Data Min. Knowl. Discov.*, 4(2-3):217–240.

Bertin, J. (1983). *Semiology of Graphics*. University of Wisconsin Press.

Blanchard, J.; Guillet, F. e Briand, H. (2003). Exploratory visualization for association rule rummaging. In *Proceedings of the KDD'2003 Workshop on Multimedia Data Mining MDM'03*, pp. 107–114.

Card, S. K. e Mackinlay, J. (1997). The structure of the information visualization design space. In *INFOVIS '97: Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97)*, pp. 92–100, Washington, DC, USA. IEEE Computer Society.

Chi, E. H. (2000). A taxonomy of visualization techniques using the data state reference model. In *INFOVIS '00: Proceedings of the IEEE Symposium on Information Vizualization 2000*, pp. 69–75, Washington, DC, USA. IEEE Computer Society.

Chin, J. P.; Diehl, V. A. e Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *CHI '88: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 213–218, New York, NY, USA. ACM Press.

Cleveland, W. S. (1985). *The Elements of Graphing Data.* Wadsworth Publ. Co.

Cleveland, W. S. (1993). *Visualizing Data.* Hobart Press.

de Souza, C. S. (2005). *The Semiotic Engineering of Human-Computer Interaction.* The MIT Press, Cambridge, MA.

Fayyad, U.; Grinstein, G. G. e Wierse, A. (2001). *Information Visualization in Data Mining and Knowledge Discovery.* Morgan Kaufmann Publishers Inc.

Fayyad, U.; Piatetsky-Shapiro, G. e Smyth, P. (1996). From data mining to knowledge discovery in databases. *Ai Magazine*, 17:37–54.

Goethals, B. e den Bussche, J. V. (1999). A priori versus a posteriori filtering of association rules. In *1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.*

Gould, J. D. e Lewis, C. (1985). Designing for usability: Key principles and what designers think. *Commun. ACM*, 28(3):300–311.

Hao, M. C.; Dayal, U.; Hsu, M.; Sprenger, T. e Gross, M. H. (2001). Visualization of directed association in e-commerce transaction data. In *VisSym '01: Proceedings of the Joint Eurographics IEEE TCVG Symposium on Visualization*, pp. 185–192.

Hofmann, H.; Siebes, A. P. J. M. e Wilhelm, A. F. X. (2000). Visualizing association rules with interactive mosaic plots. In *KDD '00: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 227–235. ACM Press.

Klemettinen, M.; Mannila, H.; Ronkainen, P.; Toivonen, H. e Verkamo, A. I. (1994). Finding interesting rules from large sets of discovered association rules. In *CIKM '94: Proceedings of the Third International Conference on Information and Knowledge Management*, pp. 401–407, New York, NY, USA. ACM Press.

Kuntz, P.; Guillet, F.; Lehn, R. e Briand, H. (2000). A user-driven process for mining association rules. In *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 483–489, London, UK. Springer-Verlag.

Liu, B.; Hsu, W. e Ma, Y. (1999). Pruning and summarizing the discovered associations. In *KDD '99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 125–134, New York, NY, USA. ACM Press.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141.

Nielsen, J. (2005). Ten usability heuristics. http://www.useit.com/papers/heuristic.

Norman, D. (1988). *The Design of Everyday Things*. MIT Press, London.

Ong, K.-H.; Ong, K.-L.; Ng, W.-K. e Lim, E.-P. (2000). Crystalclear: Active visualization of association rules.

Prates, R. O.; de Souza, C. S. e Barbosa, S. D. J. (2000). A method for evaluating the communicability of user interfaces. *Interactions*, 7(1):31–38.

Rainsford, C. e Roddick, J. (2000). Visualization of temporal interval association rules. In *IDEAL '00: In Proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning*, pp. 91–96.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages*, p. 336. IEEE Computer Society.

Shneiderman, B. (2002). Inventing discovery tools: Combining information visualization with data mining. *Information Visualization*, 1(1):5–12.

Shneiderman, B. e Plaisant, C. (2004). *Designing the User Interface: Strategies for Effective Human-Computer Interaction (4th Edition)*. Pearson Addison Wesley.

Silberschatz, A. e Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974.

Silveira, M. S.; de Souza, C. S. e Barbosa, S. D. J. (2003). A method of semiotic engineering for the online help systems construction. In *CLIHC '03: Proceedings of*

the Latin American Conference on Human-Computer Interaction, pp. 167–177, New York, NY, USA. ACM Press.

Soukup, T. e Davidson, I. (2002). *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*. John Wiley & Sons, Inc.

Tan, P.-N.; Kumar, V. e Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 32–41, New York, NY, USA. ACM Press.

Thomas, J.; Cook, K.; Crow, V.; Hetzler, B.; May, R.; McQuerry, D.; McVeety, R.; Miller, N.; Nakamura, G.; Nowell, L.; Whitney, P. e Wong, P. (1999). Human computer interaction with global information spaces - beyond data mining.

Tufte, E. R. (1986). *The Visual Display of Quantitative Information*. Graphics Press.

Tufte, E. R. (1990). *Envisioning information*. Graphics Press, Cheshire, CT, USA.

Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, Cheshire, CT, USA.

Tufte, E. R. (2005). The work of Edward Tufte and Graphics Press. http://www.edwardtufte.com.

Witten, I. H. e Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers Inc.

Wong, P. C.; Whitney, P. e Thomas, J. (1999). Visualizing association rules for text mining. In *INFOVIS '99: Proceedings of the 1999 IEEE Symposium on Information Visualization*, p. 120, Washington, DC, USA. IEEE Computer Society.

Zaki, M. e Phoophakdee, B. (2003). Mirage: A framework for mining, exploring and visualizing minimal association rules. Relatório técnico, RPI Computer Science Department Technical Report.