Visualização como suporte à extração e
exploração de regras de associação

Claudio Haruo Yamamoto

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:

Visualização como suporte à extração e exploração de regras de associação

Claudio Haruo Yamamoto

Orientadora: *Profa. Dra. Maria Cristina Ferreira de Oliveira*

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

USP – São Carlos
Fevereiro de 2009

Visualização como Suporte à Extração e Exploração de Regras de Associação

Claudio Haruo Yamamoto

Tese apresentada ao Instituto de Ciências
Matemáticas e de Computação –
ICMC/USP, como parte dos requisitos
para obtenção do título de Doutor na Área
de Ciências de Computação e Matemática
Computacional.

Orientador(a): Profa. Dra. Maria Cristina Ferreira de Oliveira

USP – São Carlos/SP
Fevereiro/2009

Dedicatória

Aos meus pais, Hayao e Rosa.

Agradecimentos

Ao final de um período de quase meia década, o mínimo que devo fazer é agradecer àqueles que foram muito importantes para mim, direta ou indiretamente, em diversos aspectos. Desta maneira, agradeço:

À minha orientadora, a professora Maria Cristina, não só pela orientação no trabalho de doutorado e pela supervisão no estágio do PAE (Programa de Aperfeiçoamento de Ensino), mas também pela convivência durante todo este tempo. Seu trabalho foi impecável.

À minha “co-orientadora”, a professora Solange Rezende, que deu o direcionamento ao trabalho quando foi preciso, numa interação com o grupo de Inteligência Computacional do ICMC/USP.

Aos professores (além da professora Solange Rezende) que participaram da banca de qualificação, o professor Mauro Biajiz – que me acompanhou durante o mestrado, sendo na ocasião meu orientador – e a professora Rosane Minghim, pelas críticas e sugestões para que o trabalho pudesse ser melhorado.

Aos professores e pesquisadores que participaram da defesa, pelas críticas e sugestões, que contribuíram para a minha formação profissional.

Ao amigos Fernando Paulovich, Márcio Almeida, Roberto Pinho, pelas sugestões e opiniões.

Aos colegas do Labic, André Maletzke, Carlos Andrés, Edson Takashi, Leonardo Almeida, Márcio Basgalupp, Ronaldo Prati, Verônica Carvalho e Walter Nagai, pelas sugestões para o trabalho.

Aos professores Alneu Lopes, Gustavo Batista, Haim Levkowitz e Maria Carolina Monard, pelas sugestões para o trabalho.

Ao Jeferson Nomelini e à ANCP (Associação Nacional de Criadores e Pesquisadores), pela colaboração intensa, que permitiu que esse trabalho pudesse ser conduzido na sua fase mais crítica, a validação.

Aos participantes dos estudos de caso.

Aos professores João Batista e Maria das Graças Nunes, o primeiro pela supervisão e a segunda pela (co-)supervisão no estágio PAE (Programa de Aperfeiçoamento de Ensino).

À Mariza, pela amizade e pela troca de experiências durante o estágio do PAE.

Aos professores Agma Traina, Dilvan Moreira, Odemir Bruno e Zazá, por participarem ativamente na minha formação, ministrando disciplinas que cursei durante o doutorado.

Aos amigos Eduardo Marmo, Leandro Henrique e Marcelo Manzato, pela convivência no período inicial do doutorado, cursando disciplinas.

Ao Anand, por auxiliar na correção de artigos em inglês.

Aos revisores dos artigos, por contribuírem com críticas e sugestões ao trabalho.

A todos os que contribuíram para o desenvolvimento deste trabalho, mas que esqueci de citar (a memória é falha).

À Magaly, pelo seu amor.

Aos meus pais e à minha irmã, por todo o suporte que me deram (e sempre me darão) para continuar a caminhada durante o doutorado, mesmo sabendo que o caminho era bastante longo e difícil.

Aos amigos André Backes, Arnaldo, Bianca, Dalcimar, Danilo, Érika, Jarbas, João Florindo, Marcelo, Marco, Simone, Ubirajara, não só pela amizade, como também pela convivência durante o doutorado.

Aos professores Adenilso, Ellen e Luciana, pela amizade.

Aos amigos e colegas que fazem ou que fizeram parte do LCAD (Laboratório de Computação de Alto Desempenho), do LABIC (Laboratório de Inteligência Computacional), do LabES (Laboratório de Engenharia de Software) e de outros laboratórios, pela troca de experiências.

À Universidade de São Paulo, incluindo o ICMC, por permitir a realização deste trabalho, e a PRPG (Pró-Reitoria de Pós-Graduação), por financiar parcialmente uma das viagens para apresentação de trabalho em conferência.

Ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), pelo suporte financeiro durante o trabalho.

A todos os funcionários do ICMC, pela excelência que permite fazer da Universidade de São Paulo uma referência mundial em ensino, pesquisa e extensão.

“Hoje entendo bem meu pai. Um homem precisa viajar. Por sua conta, não por meio de histórias, imagens, livro ou TV. Precisa viajar por si, com seus olhos e pés, para entender o que é seu. Para um dia plantar suas próprias árvores e dar-lhes valor. Conhecer o frio para desfrutar do calor. E o oposto. Sentir a distância e o desabrigo para estar bem sob o próprio teto. Um homem precisa viajar para lugares que não conhece, para quebrar essa arrogância que nos faz ver o mundo como imaginamos e não simplesmente como ele é ou pode ser. Que nos faz professores e doutores do que não vimos, quando deveríamos ser alunos, e simplesmente ir ver.”

Amyr Klink

Desde a definição do problema de obtenção de regras de associação, vários algoritmos eficientes foram introduzidos para tratá-lo. Entretanto, ainda hoje o problema apresenta várias dificuldades práticas para os mineradores, como a determinação de limiares adequados de suporte mínimo e confiança mínima, a manipulação de grandes conjuntos de regras, e a compreensão de regras (especialmente aquelas contendo muitos itens). Para tratar estes problemas, pesquisadores têm investigado a aplicação de técnicas interativas, sumarização (de conjuntos de regras) e representações visuais. Entretanto, nenhuma abordagem na qual os usuários podem entender e controlar o processo por meio da interação com o algoritmo analítico ao longo de sua execução foi introduzida. Neste trabalho, é introduzida uma abordagem interativa para extração e exploração de regras de associação que insere o usuário no processo por meio de: execução interativa do *Apriori*; seleção interativa de *itemsets* freqüentes; extração de regras baseada em *itemsets* e orientada por agrupamentos de *itemsets* similares; e exploração de regras aos pares. Para validar a abordagem, foram realizados diversos estudos, apoiados pelo Sistema I₂E, com o objetivo de: comparar a abordagem interativa, sob diversos aspectos, com uma abordagem convencional de obtenção de regras de associação; avaliar o efeito de variar alguns parâmetros do processo nos resultados finais; e mostrar a aplicação dos recursos oferecidos em situações reais e com usuários reais. Os resultados indicam que a abordagem apresentada é adequada, tanto em cenários exploratórios quanto em cenários em que há um direcionamento inicial para o

processo, à execução de certas tarefas de extração de regras de associação, pois: provém recursos capazes de evitar execuções inteiras do algoritmo antes que os resultados sejam analisados; gera conjuntos de regras mais compactos; preserva a cobertura de *itemsets*; favorece a reformulação de tarefas ou a formulação de novas tarefas; e provê meios para comparação visual de regras, aumentando o poder de análise do minerador.

Abstract

Since the definition of the association rule mining problem, many efficient algorithms have been introduced to deal with it. However, the problem still presents many practical difficulties to the miners, such as the determination of suitable minimum support and minimum confidence thresholds, manipulation of large rule sets, and comprehension of rules (specially those containing many items). In order to deal with these problems, researchers have been investigating the application of interactive techniques, summarization (of rule sets) and visual representations. Nonetheless, no approach in which users can understand and control the process through interaction with the analytical algorithm along its execution has been introduced. We introduce an interactive approach to extract and explore association rules that inserts the user into the process through: interactive execution of the *Apriori*; interactive selection of frequent itemsets; itemset-based and cluster-oriented extraction of rules; and pairwise exploration of rules. To validate the approach, several studies have been conducted, supported by the I₂E System, aiming at: comparing the interactive approach, under several aspects, with a conventional approach to obtain association rules; evaluate the effect of different execution parameters in the final results; and illustrate its application in real situations and with real users. Results of these studies indicate that the approach is adequate, both in exploratory scenarios and in scenarios in which there is an initial guidance for the process, to the execution of certain association rule extraction tasks, because: it provides resources to avoid complete algorithm executions before

results are analyzed; generates more compact rule sets for exploration; preserves rule diversity; favors the reformulation of tasks; and provides support for rule comparison, enhancing analysis capability for miners.

1	Introdução	1
1.1	Contexto e Motivação	1
1.2	Contribuições	4
1.3	Organização	6
2	Regras de Associação e Visualização de Informação	7
2.1	Considerações Iniciais	7
2.2	Mineração de Dados e Regras de Associação	8
2.3	Visualização de Informação	12
2.3.1	Tipos de Dados	13
2.3.2	Técnicas de Visualização	14
2.3.3	Técnicas de Interação e de Distorção	17
2.4	Mineração Visual de Dados	20
2.5	Considerações Finais	21
3	Apoio ao Minerador na Extração e Exploração de Regras	23
3.1	Considerações Iniciais	23
3.2	Abordagens para Diminuição do Número de Regras	23
3.3	Abordagens Interativas para a Extração de Regras	26
3.4	Abordagens Interativas com Visualização para a Exploração de Regras	28
3.5	Abordagens para Visualização durante a Obtenção de Regras	42
3.6	Considerações Finais	47
4	Abordagem Interativa para Extração e Exploração de Regras	49
4.1	Considerações Iniciais	49

4.2	Identificação do Problema – Extração de Regras Representativas	50
4.2.1	Definições	54
4.3	Metodologia Interativa para Extração de Regras Representativas	57
4.3.1	Execução Interativa do <i>Apriori</i>	58
4.3.2	Exploração Visual de <i>Itemsets</i> Frequentes	61
4.3.3	Filtragem de Regras Baseada em <i>Itemsets</i>	66
4.4	Estratégia para Exploração Direcionada de Regras	68
4.4.1	Exploração de Regras por Pares	70
4.4.2	Visualização de Medidas de Regras de Associação	72
4.5	Modelo Interativo para a Tarefa de Regras de Associação	73
4.6	Considerações Finais	75
5	Avaliação da Abordagem: Efeito dos Parâmetros no Processo	77
5.1	Considerações Iniciais	77
5.2	Avaliação da Cobertura de <i>Itemsets</i>	78
5.3	Avaliação da Política de Filtragem	82
5.3.1	Redução do número de regras	84
5.3.2	Distribuição das regras em relação ao <i>ranking</i>	86
5.4	Avaliações da Visualização de <i>Itemsets</i>	89
5.4.1	Comparação com Outras Visualizações de <i>Itemsets</i>	89
5.4.2	Escolha da Métrica de Distância	90
5.5	Avaliação do Limiar de Filtragem	95
5.6	Considerações Finais	98
6	Avaliação da Abordagem: Estudos de Caso com Usuário	99
6.1	Considerações Iniciais	99
6.2	Estudos de caso com Especialista do Domínio	100
6.2.1	Primeiro estudo de caso – PMGRN fêmeas	100
6.2.2	Segundo estudo de caso – PMGRN machos	105
6.3	Estudos de caso em Cenário Exploratório	113
6.3.1	Primeiro Estudo – Cenário Exploratório	113
6.3.2	Segundo Estudo – Tarefa Específica	119
6.4	Considerações Finais	122
7	Conclusões	123
7.1	Contribuições	124
7.2	Limitações	126
7.3	Trabalhos Futuros	127

Referências Bibliográficas	131
A Arquitetura do Sistema I ₂ E	143
B Publicações relacionadas à tese	145

Lista de Figuras

2.1	Processo de descoberta de conhecimento em bases de dados.	10
2.2	Visualização do <i>lattice</i> de uma base de dados com quatro itens.	12
2.3	Classificação das técnicas de visualização de informação.	13
2.4	Visualização de regras usando coordenadas paralelas.	15
2.5	Visualização baseada em grafos, utilizando projeção.	18
2.6	Filtragem interativa usando controles deslizantes.	19
2.7	Exemplo de ligação e seleção interativa.	20
2.8	Tipos de mineração visual de dados.	21
3.1	Visualização básica de regras usando matriz.	29
3.2	Visualização de regras, por Wong et al. (1999).	30
3.3	Visualização de regras, por Chakravarthy e Zhang (2003).	31
3.4	Visualização de regras, por Techapichetvanich e Datta (2005).	32
3.5	Visualização de regras, por Klemettinen et al. (1994).	32
3.6	Visualização de regras, por Hofmann et al. (2000).	33
3.7	Visualização de regras, por Unwin et al. (2001).	34
3.8	Protótipo CrystalClear, por Ong et al. (2002).	35
3.9	Gráfico de barras, por Kopanakis e Theodoulidis (2003).	36
3.10	Coordenadas paralelas, por Kopanakis e Theodoulidis (2003).	36
3.11	Gráfico baseado em grade, por Kopanakis e Theodoulidis (2003).	37
3.12	Visualização de regras, por Bruzzese e Buono (2004).	38
3.13	Visualização de regras, por Kuntz et al. (2000).	39
3.14	Visualização de regras, por Blanchard et al. (2003b).	40
3.15	Visualização de regras, por Couturier et al. (2007a,b).	41
3.16	Visualização de <i>itemsets</i> frequentes, por Mahanti e Alhajj (2005).	43

3.17	Visualização de <i>itemsets</i> frequentes, por Zaki e Phoophakdee (2003).	43
3.18	Visualização de <i>itemsets</i> frequentes e de regras, por Yang (2003).	44
3.19	Visualização de <i>itemsets</i> frequentes e de regras, por Ertek e Demiriz (2006).	44
3.20	Visualização de <i>itemsets</i> frequentes e de regras, por Liu e Salvendy (2006).	45
4.1	Porcentagem de regras geradas a partir dos melhores <i>itemsets</i> .	54
4.2	Interface do módulo I ₂ E.	58
4.3	Execução interativa do <i>Apriori</i> .	59
4.4	Painel de controle do Sistema I ₂ E.	59
4.5	<i>Lattice</i> e estrutura <i>trie</i> correspondente.	60
4.6	Painel de visualização do Sistema I ₂ E.	63
4.7	Escalas de cores implementadas no sistema I ₂ E.	64
4.8	Processo de filtragem de regras baseada em <i>itemsets</i> .	66
4.9	Interface do módulo RulEx.	69
4.10	Visualização em coordenadas paralelas.	73
4.11	Modelo Interativo para Extração de Regras de Associação.	74
5.1	Visualização das posições relativas das regras obtidas, por tarefa.	87
5.2	Visualização de 2- <i>itemsets</i> , base PMGRN fêmeas.	92
5.3	Visualização de 2- <i>itemsets</i> , base Orkut.	92
5.4	Visualização de 2- <i>itemsets</i> , base Corpus Artificial.	93
5.5	Visualização de 2- <i>itemsets</i> , usando diferentes métricas de distância.	94
5.6	Representação esquemática dos conjuntos de regras.	96
5.7	Análise de limiar de filtragem para a base Corpus Artificial.	97
5.8	Análise de limiar de filtragem para a base Orkut.	97
5.9	Análise de limiar de filtragem para a base PMGRN.	98
6.1	Grupos de <i>itemsets</i> extraídos da base PMGRN fêmeas.	101
6.2	Visualização de 2- <i>itemsets</i> da base PMGRN machos.	106
6.3	Comparação de regras usando o módulo RulEx.	109
6.4	Comparação de regras usando o módulo RulEx (continuação).	110
6.5	Agrupamentos de <i>itemsets</i> frequentes obtidos a partir da base Orkut.	114
6.6	Representação esquemática das associações entre os itens.	119
A.1	Arquitetura do sistema I ₂ E.	143

Lista de Tabelas

2.1	Medidas de interesse utilizadas no trabalho.	11
2.2	Agrupamentos de <i>itemsets</i> de tamanho 3.	17
3.1	Tabela comparativa das abordagens de visualização de <i>itemsets</i>	46
4.1	Atributos da base de dados <i>Iris</i>	51
4.2	Atributos da base de dados <i>Flags</i>	52
4.3	Atributos da base de dados <i>Café</i>	52
4.4	Mecanismo de ranqueamento.	53
4.5	Conjunto de transações.	55
4.6	Conjunto de regras de associação para confiança mínima 80%.	55
4.7	Conjunto reduzido de regras de associação.	56
4.8	Percentual de Redução Total Máximo de regras, uma regra por <i>itemset</i> . . .	67
5.1	Bases de dados utilizadas.	78
5.2	Cobertura de <i>itemsets</i> : base <i>Iris</i>	80
5.3	Cobertura de <i>itemsets</i> : base <i>Flags</i>	80
5.4	Cobertura de <i>itemsets</i> : base <i>Café</i>	80
5.5	Cobertura de <i>itemsets</i> após a filtragem.	81
5.6	Regras eliminadas após a filtragem de 25% das regras: base <i>Iris</i>	81
5.7	Atributos da base de dados PMGRN.	83
5.8	Características dos conjuntos de regras selecionadas, por tarefa.	84
5.9	Conjuntos de <i>itemsets</i> freqüentes obtidos, por tarefa.	85
5.10	Número de regras, em relação ao <i>powerset</i>	85
5.11	Número de regras com a abordagem I ₂ E, em relação ao <i>powerset</i>	86
5.12	Tabela comparativa das abordagens de visualização de <i>itemsets</i>	89

5.13	Atributos da tabela Comunidades.	91
5.14	Atributos da tabela MembrosComunidades.	91
5.15	Regras com suporte baixo obtidas a partir de um agrupamento de <i>itemsets</i>	94
6.1	Regras relacionadas ao ano de comercialização das fêmeas.	102
6.2	Regras relacionadas à região de origem das fêmeas comercializadas.	104
6.3	Regras relacionadas à classe Novilha para as fêmeas comercializadas.	104
6.4	Regras relacionadas à categoria PO para as fêmeas comercializadas.	104
6.5	Algumas das regras envolvendo as fazendas F, H e B.	107
6.6	Regras envolvendo uma fazenda (F, H), e dois valores de DEP.	107
6.7	Regras envolvendo uma categoria, um valor de DEP e uma fazenda.	109
6.8	Regras envolvendo ano e valor de DEP.	111
6.9	Regras envolvendo atributos de mês e de fazenda.	111
6.10	Número de regras obtidas de 2- <i>itemsets</i> , separadas por agrupamento.	115
6.11	Número de regras obtidas de 3- <i>itemsets</i> , separadas por agrupamento.	115
6.12	Regras envolvendo duas marcas de produtos.	116
6.13	Regras envolvendo três marcas de produtos.	116
6.14	Regras envolvendo um clube/liga e uma marca.	116
6.15	Regras envolvendo uma marca e uma personalidade.	116
6.16	Regras envolvendo uma marca e um <i>hobby</i>	116
6.17	Regras envolvendo uma marca e um cantor/grupo/estilo musical.	117
6.18	Regras envolvendo dois programas de TV.	117
6.19	Regras envolvendo um programa de TV e uma personalidade.	117
6.20	Regras envolvendo um programa de TV e um cantor/grupo/estilo musical.	117
6.21	Regras envolvendo um comportamento e um <i>hobby</i>	117
6.22	Regras envolvendo uma localidade e outro item.	118
6.23	Regras envolvendo um <i>hobby</i> e um cantor/grupo/estilo musical.	118
6.24	Regras envolvendo dois grupos musicais.	118
6.25	Regras envolvendo duas marcas e outro item.	118
6.26	Conjunto de regras envolvendo três marcas específicas de produtos.	119
6.27	Conjunto de regras envolvendo oito marcas específicas de produtos.	119
6.28	Conjunto de regras envolvendo diversas marcas de produtos.	120
6.29	Conjunto de regras envolvendo diversas marcas de produtos.	120

Introdução

1.1 Contexto e Motivação

A evolução da tecnologia de produção dos meios de armazenamento de dados e sua difusão contribuem para aumentar a quantidade de dados armazenados. Por outro lado, o simples armazenamento dos dados não agrega valor às instituições que os detêm, sua análise é necessária para identificar padrões interessantes de comportamento. Entretanto, a enorme quantidade de dados inviabiliza a realização desta tarefa de maneira manual. Para tornar esta tarefa viável foi criada a área de descoberta de conhecimento em bases de dados (do inglês *KDD – Knowledge Discovery in Databases*, cujo objetivo é descobrir conhecimento por meio da aplicação de algoritmos específicos para extração de padrões em dados armazenados (Fayyad et al., 1996a,b).

Uma das tarefas de mineração de dados mais conhecidas e aplicadas é a de obtenção de regras de associação. O objetivo desta tarefa é encontrar regras do tipo $A \rightarrow B$ (se A ocorre, então B também ocorre), em que A e B denotam uma ou mais entidades de interesse. Para selecionar regras interessantes, são comumente utilizados valores mínimos para duas medidas de interesse: suporte e confiança, de modo que regras que não satisfazem esses valores são descartadas. A primeira medida se refere à frequência com que A e B ocorrem juntos na base de dados, enquanto que a segunda quantifica o percentual de ocorrência de B , quando A ocorre (Agrawal et al., 1993).

Apesar de ser bastante difundida, a tarefa de obtenção de regras de associação possui diversos problemas inerentes. Alguns deles relatados na literatura são citados a seguir:

- (i) “*Analistas não sabem como deve ser o ajuste ideal de limiares mínimos*” (Ivkovic et al., 2003)
- (ii) “*(...) obtém-se bem mais regras do que [a quantidade] manipulável confortavelmente pelos humanos*” (Hofmann et al., 2000)
- (iii) “*(...) é bastante entediante para o usuário (...) achar conhecimento interessante (...) em conjuntos que podem conter centenas de milhares de regras (...)*” (Blanchard et al., 2003a)
- (iv) “*Regras são difíceis de entender.*” (Hofmann et al., 2000)
- (v) “*(...) mesmo correlações fortes entre atributos das regras obtidas não são sempre óbvias.*” (Hofmann et al., 2000)

O primeiro problema diz respeito à dificuldade de se ajustar os parâmetros iniciais, isto é, o suporte mínimo e a confiança mínima. Em geral, em um primeiro momento da mineração, adotam-se valores extremamente baixos para o suporte mínimo (por exemplo, 0,5%, 1%, 2%, 5%), e valores arbitrários para a confiança mínima. Obtido um conjunto de regras, geralmente enorme, faz-se uma análise inicial, reajustando os parâmetros para cima ou para baixo, repetidamente, até alcançar valores razoáveis. Esta atividade, é repetitiva, pouco produtiva e toma bastante tempo do usuário, mesmo para bases de dados relativamente pequenas.

Para lidar com este problema especificamente, Hidber (1999) propôs uma abordagem que permite ajustar livremente o nível de suporte durante a primeira varredura da base de dados (no máximo, são realizadas duas varreduras). Entretanto, sua abordagem armazena um superconjunto dos *itemsets* freqüentes, de modo que requer uma grande quantidade de memória. Um *itemset* freqüente é um conjunto de itens cuja freqüência é maior que a freqüência mínima (suporte mínimo) especificada. Além disso, o mecanismo não permite que o minerador se situe nem no processo de obtenção de *itemsets* freqüentes nem no processo de geração das regras. Nesta tese, é introduzida uma abordagem interativa para a obtenção de regras de associação, que permite que o usuário acompanhe o processo de obtenção de *itemsets* freqüentes, a cada passo do algoritmo, e possa reajustar o valor do suporte mínimo durante suas etapas iniciais, evitando múltiplas execuções completas do algoritmo.

O segundo e o terceiro problemas decorrem da quantidade de regras de associação obtidas, mesmo para bases de dados pequenas e com poucos itens. A explosão de combinações possíveis entre os itens gera uma enorme quantidade de *itemsets*, que por sua vez geram uma quantidade ainda maior de regras. A exploração dos conjuntos de regras resultantes pode se tornar uma atividade entediante, mesmo para conjuntos com algumas centenas de regras, dependendo do contexto em que ela é realizada, de restrições de tempo e de dinheiro, e da motivação do usuário.

Muitos trabalhos na literatura abordam estes problemas, divididos em dois grandes grupos. Em um grupo, aplica-se algum critério para reduzir o número de regras: baseado na estrutura das regras (Kryszkiewicz, 1998), (Bastide et al., 2000), (Aggarwal e Yu, 2001), (Ashrafi et al., 2004), (Zaki, 2004), (Li e Hamilton, 2004), (Pasquier et al., 2005), (Palshikar et al., 2007), (Cheng et al., 2006), (Cheng et al., 2008), em diversas medidas de interesse (Melanda e Rezende, 2004), na generalização de regras, (Domingues e Rezende, 2005), (Carvalho et al., 2007). No outro, utiliza-se algum recurso interativo para auxiliar na exploração, visando tornar o processo centrado no usuário: usando *templates* (Klemettinen et al., 1994), (Goethals e den Bussche, 2000), conceitos de consultas *online* (Aggarwal e Yu, 1998a), (Aggarwal e Yu, 2001) e agrupamento e sumarização de regras (Jorge, 2004). Os trabalhos do primeiro grupo descartam regras que possuem conhecimento representável por outra regra, mas em geral os conjuntos obtidos ainda são grandes. Os trabalhos do segundo grupo buscam aumentar a interatividade do usuário, mas não o auxiliam a identificar correlações não-óbvias entre os itens. Além disso, eles requerem que o usuário tenha diretrizes claras para a obtenção de regras, o que não acontece em cenários de busca exploratória (White et al., 2006). De modo geral, por busca exploratória entende-se a procura por informações em um cenário em que: (a) o usuário não tem certeza do que está procurando ou não tem conhecimento do contexto para formular as consultas; (b) a tarefa de pesquisa requer navegação ou exploração; ou (c) o sistema de acesso às informações é inadequado. No contexto da obtenção de regras de associação em um cenário exploratório o minerador conhece pouco sobre o tipo de informação que pode ser encontrada, portanto são geradas muitas regras e não há direção definida de exploração. Nesta tese, é introduzida uma abordagem para a seleção de regras, de acordo com o *itemset* a partir do qual ela foi gerada. Esta abordagem pode reduzir significativamente o número de regras obtidas, e produzir conjuntos de regras menores que os obtidos utilizando filtros baseados em *itemsets* frequentes fechados. Além disso, ela auxilia o usuário a inferir correlações entre itens de maneira visual.

Os dois últimos problemas citados dizem respeito à dificuldade de interpretação das regras obtidas. Considerando que foram determinados valores razoáveis das medidas de

interesse (suporte e confiança) para a filtragem, e que o número de regras obtidas é relativamente pequeno, ainda resta o problema de compreender o conhecimento obtido. Em geral, regras envolvendo poucos itens podem ser facilmente compreendidas, mas a dificuldade de interpretação aumenta proporcionalmente ao seu tamanho. Para lidar com este problema, diversas abordagens utilizam técnicas de visualização de informação para exibir regras (Hofmann et al., 2000), (Kopanakis e Theodoulidis, 2003) ou conjuntos delas usando matrizes bidimensionais (Wong et al., 1999), (Chakravarthy e Zhang, 2003), (Couturier et al., 2007a), (Couturier et al., 2007b), grades (Techapichetvanich e Datta, 2005), (Kopanakis e Theodoulidis, 2003), grafos (Klemettinen et al., 1994), (Kuntz et al., 2000), (Bruzzeze e Buono, 2004), projeções geométricas Unwin et al. (2001), (Ong et al., 2002), ou outras metáforas visuais (Blanchard et al., 2007a), (Blanchard et al., 2007b).

Estas abordagens, entretanto, limitam-se a representar visualmente regras de associação. Outras soluções representam também *itemsets* freqüentes (Mahanti e Alhajj, 2005; Zaki e Phoophakdee, 2003; Yang, 2003; Ertek e Demiriz, 2006; Liu e Salvendy, 2006). Embora estes trabalhos tentem tratar os problemas associados à obtenção de *itemsets* freqüentes, nenhum deles permite que o minerador interfira neste processo. Além disso, essas soluções, apesar de úteis quando os objetivos do usuário estão bem definidos, não dão o suporte necessário a um minerador fazendo uma busca exploratória.

1.2 Contribuições

Nesta tese, é apresentada uma abordagem interativa para obtenção de regras de associação que provê, além de uma metodologia interativa para extração de regras, uma estratégia para exploração direcionada de regras. Assim, a contribuição desta tese é esta abordagem interativa, que:

- Permite que o usuário acompanhe e direcione o processo de obtenção de *itemsets* freqüentes e de regras de associação. Em geral, nesse processo, o usuário determina os parâmetros de entrada (suporte mínimo e confiança mínima) e espera pelos resultados de maneira passiva. Acredita-se que, se o usuário assumir um papel ativo no processo, redefinindo parâmetros e indicando caminhos ao algoritmo, compreenderá melhor os resultados obtidos.
- Adota uma abordagem para seleção de regras representativas, de acordo com o *itemset* gerador. Tipicamente, o conjunto de regras resultante de uma tarefa convencional de extração de regras de associação contém muitas regras similares em termos de antecedente, conseqüente e valores para medidas de interesse. Muitas

destas regras são originadas pelos mesmos *itemsets*, de maneira que agregam pouca informação semântica ao conjunto. Tais regras não precisam necessariamente estar no conjunto de regras a serem exploradas. Garantindo que elas possam ser recuperadas se necessário, pode-se eleger um conjunto de regras representativas e apresentá-lo ao usuário, reduzindo o seu trabalho de exploração e, muitas vezes, transformando este trabalho da condição de impossível (quanto a restrições de tempo e/ou custo) para viável.

- Permite uma exploração de *itemsets* freqüentes apoiada por uma representação visual que explora o seu agrupamento em grupos similares. Este tipo de organização fornece ao usuário subsídios para inferir possíveis associações entre os itens contidos nestes grupos. A metodologia também emprega algoritmos de agrupamento para guiar a geração de regras a partir destes grupos de *itemsets*.
- Apóia a exploração de regras de associação guiada por pares, com o auxílio de uma representação visual. O pressuposto é que, ao analisar regras o fazemos por comparação por pares ou por pequenos grupos com base em características, como antecedente, conseqüente, e valores das medidas de interesse.

A abordagem interativa para obtenção de regras de associação é refletida no **Sistema I₂E**, projetado para dar suporte a ela. O sistema permite a extração interativa de regras por meio do **Módulo I₂E** e a exploração direcionada de regras por meio do **Módulo RulEx**.

As hipóteses que nortearam a realização deste trabalho são:

1. É possível facilitar (para o usuário) o processo de obtenção de regras de associação, auxiliando-o a determinar os parâmetros do algoritmo e extraindo conjuntos mais reduzidos de regras, como ponto de partida para a exploração do conhecimento.
2. É possível indicar ao usuário caminhos de exploração de regras em um cenário exploratório e aumentar a sua capacidade de interpretação dos resultados, com o apoio de representações visuais adequadas.

Para validar a abordagem, foram realizadas diversas avaliações quantitativas da abordagem, comparando-a com outras abordagens quanto à cobertura de *itemsets*, ao tamanho dos conjuntos de regras obtidos e à qualidade da visualização. Foi feita ainda uma avaliação do limiar de filtragem, com o objetivo de auxiliar na sua determinação, para bases de dados com características distintas. Foram conduzidos ainda estudos de caso com usuário: com especialista, quando há um direcionamento inicial do processo, e em cenários totalmente exploratórios, em que não há nenhum direcionamento.

1.3 Organização

Esta tese está organizada da seguinte maneira. No **Capítulo 2**, é introduzida a tarefa de obtenção de regras de associação e fundamentos de visualização de informação. No **Capítulo 3**, são discutidas abordagens encontradas na literatura que tratam os problemas discutidos anteriormente e suas limitações. No **Capítulo 4**, é apresentada a abordagem interativa proposta para obtenção de regras de associação baseada em *itemsets* e orientada por agrupamentos e para exploração visual de regras de associação por comparação em pares. Nesse capítulo, é apresentado ainda o sistema I₂E, construído para apoiar a abordagem proposta: são discutidos os aspectos principais e as decisões tomadas durante o projeto. No **Capítulo 5**, são descritas as avaliações quantitativas da abordagem interativa, comparando-a com outras abordagens, e observando o efeito de variar alguns parâmetros de execução nos resultados do processo. No **Capítulo 6**, são descritos estudos de caso conduzidos com usuários reais e experimentais, buscando verificar outros aspectos da metodologia que não podem ser mensurados junto a potenciais usuários. Por fim, no **Capítulo 7**, são apresentadas as conclusões do trabalho e direções para trabalhos futuros.

Regras de Associação e Visualização de Informação

2.1 Considerações Iniciais

Nas últimas décadas, a mineração de dados e a visualização de informações se consolidaram como áreas de relevância dentro da computação. A primeira surgiu como o objetivo de obter padrões contidos nos dados, seja para descrevê-los, seja para prever o comportamento de novos dados. A segunda, com o objetivo de representar dados abstratos com o uso de metáforas visuais para apoiar a exploração dos mesmos. Mais recentemente, estas duas áreas vêm sendo combinadas em uma nova área que vem sendo chamada de mineração visual de dados. Esta última tem como objetivo obter padrões contidos nos dados, apoiado em metáforas visuais para auxílio no processo.

Neste capítulo, serão apresentados conceitos básicos de mineração de dados, especificamente da tarefa de obtenção de regras de associação, de visualização de informação e de mineração visual de dados. Depois disso, serão apresentados trabalhos relevantes encontrados na literatura que tentam tratar os problemas discutidos no capítulo anterior, voltados para a tarefa de obtenção de regras de associação. Estes trabalhos englobam abordagens para diminuição do número de regras obtidas, para exploração de conjuntos de regras, ambas sem o uso de recursos visuais e abordagens que utilizam representações visuais para apoiar a tarefa de obtenção de regras de associação.

2.2 Mineração de Dados e Regras de Associação

O processo de descoberta de conhecimento em bases de dados, conhecido por *KDD* (*Knowledge Discovery in Databases*), é interativo e iterativo, envolvendo uma série de etapas, nas quais importantes decisões são feitas pelo usuário (Fayyad et al., 1996a,b). O processo é interativo porque envolve diretamente o usuário no tratamento e processamento dos dados; e é também iterativo porque pode envolver diversas iterações das etapas.

Neste trabalho, é considerada a descrição do processo de descoberta de conhecimento apresentada na **Figura 2.1** (Rezende et al., 2003), composta por três grandes etapas: o **pré-processamento**, a **extração de padrões** e o **pós-processamento**. Estas etapas são precedidas pela **identificação do problema** e sucedidas pela **utilização do conhecimento** obtido no processo. Na prática, o processo não é bem comportado, no sentido em que ele não segue linearmente as três etapas e finaliza, de modo que podem ocorrer múltiplos ciclos envolvendo as três etapas, como ilustrado na figura. Um dos objetivos deste trabalho é, justamente, reduzir o número de ciclos envolvendo as três etapas em questão, diminuindo o esforço necessário para realizar a extração de padrões.

Na **identificação do problema**, deve ser realizado um estudo de domínio do problema, em que são identificados os objetivos a serem atingidos com a mineração de dados. Pessoas com um bom nível de conhecimento sobre o domínio são necessárias neste estágio.

O **pré-processamento** pode envolver uma série de atividades, tais como a obtenção e unificação, limpeza e a redução do volume dos dados. A unificação é necessária pelo fato de os dados geralmente serem produzidos em formatos diferentes e por fontes distintas. A limpeza dos dados consiste em dar um tratamento adequado a dados errados ou incompletos, como valores inválidos ou com granularidade incorreta, causados, por exemplo, por erros de digitação, de coleta, etc. Por fim, pode ser necessário reduzir o volume dos dados, devido a limitações de memória ou de tempo de processamento. Isso pode envolver a redução do número de exemplos, por meio de amostragem significativa; do número de atributos, por meio da eliminação de atributos menos significativos; e do número de valores de um atributo, geralmente substituindo valores de atributo contínuo por intervalos discretos; ou por suavização, na qual os valores de um atributo são substituídos por um valor numérico representativo, como a média, a mediana ou mesmo valores extremos.

A **extração de padrões** é a etapa principal do processo de descoberta de conhecimento. Nela, é feita a escolha da função a que se destina, se preditiva ou descritiva, a partir da qual, define-se alguma tarefa de mineração. A **predição** envolve o uso de variáveis ou de campos da base de dados para prever valores desconhecidos ou futuros de variáveis de interesse. Por outro lado, a **descrição** busca identificar a presença de

padrões interpretáveis pelos humanos (Fayyad et al., 1996a). Daí, escolhe-se o algoritmo a ser utilizado. Pode ser necessário aplicar transformações aos dados de maneira que eles possam ser submetidos à mineração. Por fim, é executada a tarefa propriamente dita, a extração dos padrões, objeto principal do processo. São exemplos de tarefas: classificação, regras de associação, agrupamento, regressão, séries temporais. A **classificação** é o aprendizado de uma função que mapeia um item de dado em uma de várias classes. **Regras de associação** permitem relacionar a ocorrência de um determinado conjunto de itens com a ocorrência de outro conjunto de itens. Esta tarefa é abordada em detalhes mais adiante nesta mesma seção. O **agrupamento** (*clustering*) é uma tarefa descritiva na qual se busca identificar um conjunto finito de categorias, ou agrupamentos (*clusters*), para descrever os dados. A **regressão** é o aprendizado de uma função que mapeia um item de dado a um valor real de variável de predição. Em **séries temporais**, padrões de comportamento são obtidos tanto para descrever o comportamento dos dados como para prever comportamento futuro.

Após a extração de padrões, pode-se analisar os resultados obtidos, em uma etapa de **pós-processamento**. Um tipo de pós-processamento é a filtragem dos padrões obtidos, atividade interessante quando a quantidade de padrões é bastante elevada, o que dificulta o entendimento. Pode-se ainda interpretar e explicar os resultados obtidos, atividade realizada em geral quando o conhecimento é utilizado pelo usuário final na tomada de decisão. Neste caso, o conhecimento pode ser tratado de maneira a torná-lo compreensível ao usuário, por meio de sumarização, visualização ou combinação com conhecimento prévio sobre o domínio. Também costuma-se avaliar o conhecimento obtido, de maneira a determinar a sua qualidade, utilizando critérios específicos para cada tarefa. Por fim, pode-se integrar o conhecimento com o resultado de extrações prévias de conhecimento.

Finalmente, tem-se a etapa de **utilização do conhecimento**, que pode ocorrer por meio da incorporação a um sistema inteligente, da utilização direta pelo usuário final em alguma atividade de tomada de decisão na sua instituição ou de relatórios a serem repassados às pessoas interessadas.

Neste trabalho, será abordado o problema de obtenção de regras de associação, que pode ser definido da seguinte maneira. Seja $\mathcal{I} = \{I_1, I_2, \dots, I_m\}$ um conjunto de atributos binários, chamados itens. Seja T uma base de dados de transações. Cada transação $t \in T$ é representada como um vetor binário, com $t[k] = 1$ se t inclui o item I_k , e $t[k] = 0$ caso contrário. Existe uma tupla na base de dados para cada transação. Seja X um conjunto de itens em $\mathcal{I}$. Diz-se que uma transação satisfaz X se, para todo item I_k em X , $t[k] = 1$. O **suporte** de um conjunto de itens X , denominado *itemset* é dado pelo número de transações que o satisfazem. A partir de um *itemset* X , podem ser obtidas **regras**

Figura 2.1: Processo de descoberta de conhecimento em bases de dados.

de associação, que são implicações do tipo $Y \rightarrow (X - Y)$, tal que $Y \subset X$ e $Y \neq \emptyset$. A **confiança** de uma regra é dada pela relação entre o suporte da regra (suporte de todos os itens da regra) e o suporte do antecedente.

O problema de obter regras de associação pode ser dividido em dois problemas menores. Por ser complexo, o primeiro problema é bastante estudado, e existem diversos algoritmos na literatura para resolvê-lo. Já o segundo é considerado simples, e em geral é apenas indicado.

Problema 1 – Obtenção de *itemsets* freqüentes (*FI – Frequent Itemsets*). Gerar todas as combinações X de itens pertencentes a $\mathcal{I}$ que têm suporte transacional acima de um certo valor limite especificado, chamado **suporte mínimo**. Chamar estas combinações de ***itemsets* freqüentes**, e todas as outras combinações de ***itemsets* não freqüentes**.

Problema 2 – Geração de regras de associação. A partir de cada *itemset* freqüente X , gerar todas as regras (no máximo $2^k - 2$), onde k é o número de itens de X . Determinar todos os subconjuntos próprios de itens $S \subset X$, excluindo o conjunto

vazio, para serem os antecedentes¹ das regras. Os conseqüentes são determinados por $(X - S)$. Se o valor para a confiança da regra é maior ou igual a um certo valor limite especificado, chamado **confiança mínima**, a regra é considerada. Caso contrário, a regra é descartada.

O primeiro problema produz um conjunto com todos os *itemsets* freqüentes. Para representar este conjunto, pode-se utilizar o conjunto contendo apenas os chamados ***itemsets* freqüentes maximais (MFI – Maximal Frequent Itemsets)** aqueles que não são subconjunto de nenhum outro *itemset* freqüente. Um conjunto de *itemsets* freqüentes maximais pode ser utilizado para representar *itemsets* freqüentes de maneira compacta. Entretanto, como somente o suporte dos *itemsets* freqüentes maximais é guardado, perde-se informação relativa ao suporte dos outros *itemsets* freqüentes.

Uma maneira de representar conjuntos de *itemsets* freqüentes sem perder informação do suporte é utilizar os ***itemsets* freqüentes fechados (FCI – Frequent Closed Itemsets)**. Um *itemset* freqüente X é chamado *itemset* freqüente fechado se não existe nenhum superconjunto próprio $Y \supset X$, tal que $\text{suporte}(X) = \text{suporte}(Y)$. Apesar de os conjuntos de *FCI* serem normalmente maiores que os de *MFI*, eles preservam informação de suporte.

Além do suporte e da confiança, existem diversas medidas de interesse para regras de associação. Neste trabalho, são utilizadas oito medidas de interesse com características diferentes para a composição de *rankings*, discutidos na **Seção 4.2**. O critério para escolha destas medidas teve como base o trabalho de Tan et al. (2004), em que as principais medidas de interesse foram separadas em grupos com características semelhantes. Para este trabalho, foi escolhida uma medida representativa de cada um dos grupos. As medidas de interesse escolhidas são listadas na **Tabela 2.1**.

Tabela 2.1: Medidas de interesse utilizadas no trabalho.

Medida	Intervalo	Equação
Lift	$[0 \dots 1 \dots \infty]$	$LF = \frac{P(AB)}{P(A)P(B)}$
Leverage	$[-0.25 \dots 0 \dots 0.25]$	$LEV = P(AB) - P(A)P(B)$
Certainty Factor	$[-1 \dots 0 \dots 1]$	$F = \frac{P(B A) - P(B)}{1 - P(B)}$
Jaccard	$[0 \dots 1]$	$\zeta = \frac{P(AB)}{P(A) + P(B) - P(AB)}$
Gini Index	$[0 \dots 1]$	$G = P(A)[P(B A)^2 + P(B A)] + P(A)[P(B A)^2 + P(B A)^2] - P(B)^2 - P(B)^2$
Yule's Q	$[-1 \dots 0 \dots 1]$	$Q = \frac{P(AB)P(AB) - P(AB)P(AB)}{P(AB)P(AB) + P(AB)P(AB)}$

A visualização do *lattice* ilustra as relações entre os *itemsets* de uma base de transações. A visualização, baseada em grafos, consiste em mapear em cada nó um *itemset* freqüente,

¹Alternativamente, chama-se o antecedente de *LHS* (*left-hand side*) e o conseqüente de *RHS* (*right-hand-side*).

em camadas, de maneira que cada camada inclua apenas *itemsets* frequentes de mesmo tamanho. Cada arco liga dois *itemsets*, de modo que um deles seja um subconjunto do outro e que ele seja de tamanho imediatamente inferior. Um exemplo de *lattice* é mostrado na **Figura 2.2**, para uma base de dados com quatro itens (A, B, C e D).

Figura 2.2: Visualização do *lattice* de uma base de dados com quatro itens.

2.3 Visualização de Informação

O objetivo da visualização é criar modelos gráficos e representações visuais a partir de dados, com o objetivo de apoiar a interação do usuário com estes dados, facilitando a exploração e a aquisição de informações úteis (de Oliveira e Levkowitz, 2003). Pode-se identificar duas linhas em visualização, a Visualização Científica e a Visualização de Informação, embora esta diferenciação nem sempre seja adequada ou necessária. A grosso modo, na Visualização Científica, os modelos gráficos são, tipicamente, construídos a partir de dados que representam objetos ou conceitos associados a fenômenos do mundo físico, definidas em espaços unidimensionais, bidimensionais ou tridimensionais. Eventualmente, pode-se fazer uso do espaço 4D considerando também a dimensão temporal. Por outro lado, os modelos gráficos podem representar conceitos abstratos ou relacionamentos que não têm, necessariamente, atributos espaciais, como textos ou registros de acessos a páginas de um portal da *Web*. Estes modelos gráficos são, tipicamente, objeto de estudo da **Visualização de Informação (IV – *Information Visualization*)**, ou Visualização Exploratória. Pelo fato de representar dados que não têm uma representação espacial inerente, o desafio da Visualização de Informação é criar novas metáforas visuais para representar informações e propor maneiras de manipular estas metáforas de maneira

a transmitir o sentido da informação. Em decorrência deste desafio, as representações visuais geradas na Visualização de Informação são menos intuitivas e, em geral, requerem algum treinamento dos usuários para que estes possam utilizá-las efetivamente.

Diversas técnicas de IV são descritas na literatura. Keim e Kriegel (1996) propõem uma taxonomia de técnicas com seis categorias: **projeção geométrica**, **iconográficas**, **baseadas em pixels**, **hierárquicas**, **baseadas em grafos** e **híbridas**, além das técnicas clássicas (gráficos de barras, de linhas, x-y-z). Posteriormente, Keim (2002) estende esta classificação, considerando dois critérios adicionais de categorização (**Figura 2.3**): o tipo dos dados a serem visualizados e a técnica de interação e de distorção utilizada. Nesta classificação, estes dois critérios de categorização são tratados como ortogonais, de modo que qualquer técnica de visualização pode ser combinada com qualquer técnica de interação e de distorção e com qualquer tipo de dados, como mostra a **Figura 2.3**. Entretanto, sabe-se que, para cada contexto, podem existir uma ou mais técnicas mais apropriadas, e este julgamento em geral é feito de maneira *ad hoc*. Neste trabalho são descritas com mais detalhes as técnicas de projeção geométrica e as baseadas em grafos, e as técnicas de filtragem interativa e de seleção e ligação, pois são empregadas na abordagem.

Figura 2.3: Classificação das técnicas de visualização de informação.

2.3.1 Tipos de Dados

Keim (2002) classificou os dados em: unidimensionais; bidimensionais; multidimensionais; texto e hipertexto; hierarquias e grafos; e algoritmos e *software*. Dados unidimensionais são aqueles que possuem apenas uma dimensão densa. Dados temporais, como preço de ações em um intervalo de tempo, são exemplos de dados unidimensionais. Dados bidimensionais são aqueles que possuem duas dimensões distintas, por exemplo, dados

geográficos com informações sobre latitude e longitude. Dados multidimensionais são aqueles que possuem mais de três dimensões, de modo que não podem ser representados diretamente em uma simples visualização de duas ou três dimensões. Dados de tabelas relacionais são exemplos de dados multidimensionais. Textos e hipertextos diferem dos anteriores porque não podem ser representados por atributos numéricos e, portanto não podem ser visualizados eficientemente com técnicas convencionais. Hierarquias e grafos são tipos de dados que podem representar relacionamentos entre conceitos. Por exemplo, pode-se representar o organograma de uma empresa, os relacionamentos entre as pessoas, os diretórios de um sistema de arquivos etc. Por fim, algoritmos e *software* podem ser visualizados como suporte ao desenvolvimento pela ajuda no entendimento de algoritmos. Por exemplo, pode-se utilizar para compreender o fluxo de um trecho de código, para visualizar erros no código como um todo etc.

2.3.2 Técnicas de Visualização

As técnicas de **projeção geométrica** buscam encontrar projeções “interessantes” de dados definidos em espaços multidimensionais, as quais possam ser exibidas na tela do computador. São exemplos desta categoria: os gráficos de dispersão (*scatterplots*) (Cleveland, 1993); Coordenadas Paralelas (*Parallel Coordinates*) (Inselberg e Dimsdale, 1990; Inselberg, 1997); Coordenadas Paralelas Circulares (Hoffman, 1999); *Radial Coordinate Visualization (RadViz)* (Hoffman, 1999); *Prosection views* (Furnas e Buja, 1994; Spence et al., 1995); e *Hyperslice* (van Wijk e van Liere, 1993).

Dentre as técnicas citadas, uma das mais conhecidas é Coordenadas Paralelas. A idéia desta técnica é mapear dados k -dimensionais ou objetos do espaço em uma visualização em 2D, desenhando eixos paralelos igualmente espaçados. Um exemplo é mostrado na **Figura 2.4**. Neste exemplo, cada um dos cinco eixos corresponde a uma medida de interesse – suporte, confiança, *lift*, *leverage* e *certainty factor* – de modo que o menor valor é representado no extremo inferior, enquanto que o maior valor é representado no extremo superior. Duas regras são mostradas: *Traquinas* $\rightarrow$ *TheSimpsons* (regra selecionada, com suporte 18, confiança 0,295, *lift* 4,646, *leverage* 0,009 e *certainty factor* 0,247), representada pela polilinha em vermelho, e *TheSimpsons* $\rightarrow$ *Traquinas* (regra de referência, com valores 18, 0,176, 4,646, 0,009, 0,144), representada pela polilinha em azul. A representação visual favorece a comparação rápida entre os valores das duas regras – a polilinha em vermelho contém valores maiores para as medidas de interesse.

As técnicas **iconográficas** mapeiam cada item de dados multidimensional em um ícone, cujas características visuais são manipuladas de maneira a refletir os valores que representam. Aqui, podem ser citadas as Faces de *Chernoff* (Chernoff, 1973); *Stick Figu-*

Figura 2.4: Visualização de regras usando coordenadas paralelas.

res (Pickett e Grinstein, 1988); *Shape Coding* (Beddow, 1990); *Color Icons* (Levkowitz, 1991; Keim e Kriegel, 1994); *Star Icons* (Ward, 1994); *Sound Icons* (Levkowitz et al., 1995); e *Tile Bars* (Hearst, 1995).

Nas técnicas **baseadas em pixels**, cada item de dado é mapeado para um pixel colorido, e os valores de cada atributo são exibidos em janelas separadas. Keim e Kriegel (1996) e Keim (2000), introduziram esta estratégia e diversas variações, como o *Recursive Pattern* (Keim et al., 1995), *Segmentos Circulares* (Ankerst et al., 1996) e *Pixel Bar Charts* (Keim et al., 2001).

Em todas estas variações, existem diversas questões a serem definidas para a elaboração de uma boa visualização. Primeiramente, deve-se utilizar um bom mapeamento dos valores dos dados em cores. A segunda questão a ser considerada é a maneira como os pixels são organizados dentro das janelas. Por exemplo, em (Keim, 2000), são mostradas as distribuições de Peano-Hilbert e de Morton. Outra questão é o formato da janela a ser adotado. Por exemplo, em *Segmentos Circulares* (Ankerst et al., 1996), a janela é circular, enquanto que em (Keim et al., 1995), ela é quadrada. Por fim, deve-se considerar a seqüência de exibição das dimensões, pois ela afeta a qualidade do resultado.

As técnicas **hierárquicas** subdividem o espaço multidimensional e apresentam os sub-espacos obtidos a cada divisão de maneira hierárquica. Aqui, podem ser citadas as técnicas: *Dimensional stacking* (LeBlanc et al., 1990); *Cone Trees* (Robertson et al., 1991); *Treemaps* (Johnson e Shneiderman, 1991; Shneiderman, 1992); e *Worlds-within-Worlds* (também conhecido como *n-Vision*) (Feiner e Beshers, 1990; Beshers e Feiner, 1993).

As técnicas **baseadas em grafos** geram representações de grafos usando algoritmos de *layout* específicos. Apesar de o estudo de visualização de grafos estar, de certa maneira,

relacionado ao problema de desenho de grafos (*graph drawing*), algumas questões-chave aparecem especificamente no que diz respeito à visualização. Enquanto em desenho de grafos a preocupação é calcular o posicionamento de nós e arestas de modo que estes obedeçam a determinadas regras estéticas, em visualização a maior preocupação é representar elementos de dados e seus relacionamentos.

Neste trabalho, serão utilizadas técnicas de projeção para posicionar os vértices dos grafos que representam *itemsets*. As técnicas de projeção podem ser divididas em dois grandes grupos: técnicas lineares e as técnicas não-lineares. As primeiras derivam combinações lineares dos atributos de dados, mapeando-os para um novo sistema de coordenadas ortogonais de poucas dimensões. Estas técnicas, apesar de terem um bom desempenho em geral, falham ao tentar representar conjuntos de dados com padrões não usuais, como agrupamentos. As técnicas não-lineares, por outro lado, tratam o problema como um processo de otimização, em que a idéia é diminuir a perda de informação dada por uma função de distância. Estas técnicas, como utilizam medidas de dissimilaridade para representar as relações de distância entre os dados, conseguem melhor representar agrupamentos, além de obterem um bom compromisso entre escalabilidade e precisão.

Recentemente, foi desenvolvida a técnica não-linear *IDMAP* (*Interactive Document Map*) (Minghim et al., 2006), que utiliza uma estratégia de *Force-Directed Placement* (*FDP*) (Tejada et al., 2003), combinada com uma técnica de redução de dimensionalidade chamada *FastMap* (Faloutsos e Lin, 1995). No trabalho de Paulovich (2008), outras duas técnicas não-lineares foram desenvolvidas: *ProjClus* (*Projection by Clustering*) (Paulovich e Minghim, 2006) e *LSP* (*Least-Squares Projection*) (Paulovich et al., 2008). A primeira emprega algoritmos de agrupamento e, sobre cada agrupamento, aplica a *IDMAP* para posicionar os seus pontos, objetivando reduzir a complexidade computacional desta última. A segunda, por sua vez, obtém pontos de controle, utilizados para determinar sistemas lineares que, resolvidos utilizando a técnica *least squares*, fornecem o posicionamento dos demais pontos vizinhos a eles. Dentre estas técnicas, a *IDMAP* é mais precisa, porém menos eficiente que as demais (*ProjClus* e *LSP*). Além disso, estas duas últimas são particularmente mais indicadas para conjuntos de pontos distribuídos em agrupamentos bem definidos.

Na **Figura 2.5** é mostrado um exemplo de visualização de *itemsets* freqüentes baseada em grafos. Neste exemplo, é utilizada a base Orkut, cujos registros descrevem informações relativas a membros participantes de diversas comunidades do Orkut (por exemplo, “Peugeot 206 Brasil”, “FIAT – Movidos pela Paixão” etc.). Neste caso, as transações informam as comunidades às quais um determinado membro pertence. Na visualização em questão, é representado um conjunto de *itemsets* de tamanho 3. Nesta representação, cada círculo

representa um *itemset* freqüente, e cada linha ligando dois círculos representa itens em comum entre os *itemsets* freqüentes em questão. O posicionamento dos objetos é dado por meio da técnica *IDMAP* sobre os itens contidos nos *itemsets*. Podem ser identificados na representação visual sete grandes grupos de *itemsets* (foi usado o algoritmo *k*-médias para determinar os grupos, com $k = \sqrt{n}$, onde n é o número de atributos – comunidades) que possuem comunidades com maior grau de similaridade entre si. Pelo conteúdo, mostrado na **Tabela 2.2**, é possível perceber que os agrupamentos mais próximos têm um maior número de comunidades em comum, como é o caso do amarelo com o preto, do preto com o verde, do verde com o rosa, e do rosa com o azul esverdeado. Os grupos branco e azul, por outro lado, ficam mais isolados em relação aos outros. Isto mostra a capacidade de preservar as relações de similaridade/dissimilaridade das técnicas de projeção não-lineares.

Tabela 2.2: Agrupamentos de *itemsets* de tamanho 3.

Grupo (Cor)	Itens predominantes no agrupamento
branco	<i>Peugeot 206 Brasil</i> (6 ocorrências), <i>FIAT - Movidos pela Paixão</i> (6) <i>Volkswagen do Brasil</i> (6), <i>Volkswagen Gol</i> (6) e <i>Volkswagen Voyage</i> (6)
azul	<i>Se meus olhos tirassem fotos..</i> (6) e <i>Já desci escada pelo corrimão!</i> (6)
amarelo	<i>Ayrton Senna</i> (14) e <i>Mamonas Assassinas</i> (9)
preto	<i>A gente se F*** mas se DIVERTE</i> (32) e <i>Mamonas Assassinas</i> (8)
verde	<i>Adidas</i> (23), <i>A gente se F*** mas se DIVERTE</i> (8), <i>Puma</i> (7) <i>F****!</i> (6) e <i>Tuning</i> (5)
rosa	<i>Puma</i> (15), <i>Adidas</i> (15), <i>Nike</i> (13) e <i>Tuning</i> (7)
azul esverdeado	<i>Eu já Fui um Espermatozóide</i> (4), <i>Billabong</i> (3), <i>Quiksilver</i> (3) <i>F * * * - S E</i> (3) e <i>Tuning</i> (3)

Por fim, existem as técnicas **híbridas**, resultado de combinações de características de mais de uma categoria. Pode-se mencionar, por exemplo, a técnica *Hierarchical Pixel Bar Charts* (Keim et al., 2002), uma extensão da técnica *Pixel Bar Charts* que permite visualizar dados hierárquicos. A idéia básica é mostrar, em gráficos de barra, os valores de dados um a um por meio de *pixels*, ao invés de agregá-los.

2.3.3 Técnicas de Interação e de Distorção

O objetivo principal de uma técnica de interação é dar ao analista recursos para interagir diretamente com a visualização. Uma técnica de distorção, por outro lado, fornece meios de focalizar em detalhes, preservando a visão global. Para efeito de classificação, Keim (2002) dividiu as técnicas em cinco tipos distintos: projeções dinâmicas; filtragem interativa; *zoom* interativo; distorção interativa; e ligação e seleção interativas.

A idéia básica de **projeções dinâmicas** é mudar dinamicamente as projeções para melhor explorar conjuntos de dados multidimensionais. Esta técnica permite que projeções bidimensionais interessantes sejam obtidas em gráficos de dispersão simples. Uma

Figura 2.5: Visualização baseada em grafos, utilizando projeção.

desvantagem é que o número de projeções cresce exponencialmente à medida que o número de dimensões cresce, o que torna inviável a exploração de conjuntos de dados de dimensionalidade muito alta. O *Grand Tour* e o *XGobi* (Asimov, 1985) são exemplos de sistemas que fornecem a técnica de projeções dinâmicas.

Por outro lado, o objetivo da **filtragem interativa** é selecionar interativamente subconjuntos de dados de interesse, o que pode ser feito por meio de seleções ou por meio de consultas. São exemplos as consultas dinâmicas (*dynamic queries*) (Shneiderman, 1994), fazendo uso de controles deslizantes (*sliders*) e de botões de seleção, como os de rádio (*radio box*) e de checagem (*check box*). Os controles deslizantes permitem selecionar intervalos de valores; os botões de rádio permitem selecionar uma opção das apresentadas e os botões de checagem permitem selecionar ou não cada opção das apresentadas.

A **Figura 2.6** exemplifica o uso de filtragem interativa. Neste exemplo, um conjunto de regras de associação é mostrada na tabela, na parte superior à esquerda, e a idéia é ocultar as regras que não satisfazem ao valor mínimo indicado em cada um dos controles deslizantes. Na figura, o controle de confiança (*confidence*), que vai de 0,39 a 0,91, indica que apenas as regras com confiança maior ou igual a 0,59 são mostradas na tabela, sendo todas as demais ocultadas.

Figura 2.6: Filtragem interativa usando controles deslizantes.

No **zoom interativo**, a idéia é mostrar, por exemplo, dados tabulares, em diferentes resoluções, permitindo visualizar uma grande quantidade de dados ao mesmo tempo que o foco do usuário se mantém nos dados de interesse. Assim, é possível ter uma visão geral dos dados e, por meio de seleção de porções dos dados, focalizar elementos de interesse, facilitando a exploração.

Na **distorção interativa**, os dados são exibidos em diferentes níveis de resolução. Deste modo, a visualização é interativamente distorcida, de maneira que parte da representação é exibida como se uma lente de aumento estivesse sobre ela, enquanto que o restante é exibido de maneira convencional. *Perspective Wall* (Mackinlay et al., 1991), *Graphical Fisheye Views* (Sarkar e Brown, 1994) e *Hyperbolic Visualization* (Lamping et al., 1995) são alguns exemplos de técnicas que utilizam este princípio.

Por fim, a **ligação e seleção interativa** permitem coordenar a interação com visualizações distintas. Por meio desta técnica, pode-se atuar sobre um mesmo objeto em várias visualizações, de modo que quando se realiza uma operação sobre um ou mais objetos mapeados, ela se reflete nas demais. O objetivo é tornar a tarefa de exploração mais eficiente, simplificando a interação e favorecendo a identificação de correlações.

Na **Figura 2.7**, é mostrado um exemplo de ligação e seleção interativa entre a visualização baseada em grafo (representando *itemsets* frequentes) e um histograma (representando a distribuição de suporte dos *itemsets*), ambos à direita, e a lista de itens, à esquerda. Uma seleção na visualização de *itemsets* frequentes é refletida tanto no histograma, no qual são selecionadas as colunas representando os *itemsets*, quanto na lista de itens, na qual são selecionados os itens contidos nos *itemsets* selecionados.

Figura 2.7: Exemplo de ligação e seleção interativa.

2.4 Mineração Visual de Dados

Mais recentemente, a área de **Mineração Visual de Dados**, que combina a motivação da mineração de dados e a utilização de ferramentas analíticas visuais vem ganhando espaço. Segundo Ankerst (2000), a interação entre mineração de dados e técnicas de visualização pode ser classificada em três tipos: **visualização de modelos de mineração**, **mineração visual de dados** e **visualização exploratória**. Eles são ilustrados, respectivamente, nas **Figuras 2.8(a)**, **2.8(b)** e **2.8(c)**.

No primeiro tipo, visualiza-se os resultados obtidos a partir da aplicação de algoritmos de mineração de dados. Este tipo é o mais adotado na tarefa de regras de associação, segundo o levantamento bibliográfico realizado. No segundo tipo, visualiza-se resultados intermediários durante a aplicação de algoritmos de mineração de dados. Desta maneira, permite-se que o usuário/minerador acompanhe o desenrolar do processo, e o direcione

Figura 2.8: Tipos de mineração visual de dados.

segundo os seus objetivos. As abordagens mais recentes de aplicação de visualização na tarefa de regras de associação se enquadram neste segundo tipo, revelando uma tendência. Por fim, o terceiro tipo refere-se a um processo de exploração visual dos dados brutos, sem a aplicação de algoritmos de mineração.

2.5 Considerações Finais

Neste capítulo, foram introduzidos os conceitos básicos e técnicas fundamentais da extração de regras de associação e visualização de informação.

No próximo capítulo, são discutidas abordagens encontradas na literatura que tentam tratar os problemas levantados na mineração de regras de associação, bem como as suas vantagens e limitações.

Apoio ao Minerador na Extração e Exploração de Regras

3.1 Considerações Iniciais

Nesta seção, são apresentados diversos trabalhos que tentam tratar os problemas discutidos no capítulo anterior. Eles são divididos em quatro grandes grupos. O primeiro engloba as abordagens que lidam com a explosão de regras de associação. O segundo engloba as abordagens que adotam algum tipo de recurso para aumentar a interatividade do processo para auxiliar o usuário na exploração. O terceiro reúne abordagens que trazem representações visuais para regras de associação. Finalmente, o último grupo contém abordagens que recorrem tanto à visualização das regras de associação quanto dos *itemsets* freqüentes, numa tendência de fortalecer o papel das representações visuais no processo de mineração.

3.2 Abordagens para Diminuição do Número de Regras

Kryszkiewicz (1998) definiu regras redundantes como sendo aquelas que têm antecedente mínimo e conseqüente máximo. Ele definiu um operador, dado por $C(X \rightarrow Y) = \{X \cup Z \rightarrow V | Z, V \subseteq Y \text{ e } Z \cap V = \emptyset \text{ e } V \neq \emptyset\}$. Então, dado o conjunto inicial

de regras de associação AR (obtido com um suporte mínimo e uma confiança mínima determinados), o conjunto de regras de associação representativas RR é definido como sendo $RR(s, c) = \{r \in AR(s, c) | \nexists r' \in AR(s, c), r' \neq r \text{ e } r \in C(r')\}$. Deste conjunto, é extraído um subconjunto de regras do tipo antecedente mínimo, conseqüente máximo (*Minimum Condition Maximum Consequence Association Rules*), denominado MMR . Este conjunto é dado por $MMR(s, c) = \{r : (X \rightarrow Y) \in AR(s, c) | \nexists r' : (X' \rightarrow Y') \in AR(s, c), r' \neq r \text{ e } X' \subseteq X \text{ e } Y' \supseteq Y\}$. Por exemplo, a regra $B \rightarrow CDE$, que tem antecedente mínimo e conseqüente máximo, é representativa de $B \rightarrow CD$, $B \rightarrow CE$, $B \rightarrow DE$, $B \rightarrow C$, $B \rightarrow D$, $B \rightarrow E$, $BC \rightarrow DE$, $BC \rightarrow D$, $BC \rightarrow E$, $BD \rightarrow CE$, $BD \rightarrow C$, $BD \rightarrow E$, $BE \rightarrow CD$, $BE \rightarrow C$, $BE \rightarrow D$, $BCD \rightarrow E$, $BCE \rightarrow D$ e $BDE \rightarrow C$.

Bastide et al. (2000) e Aggarwal e Yu (2001) seguem a mesma linha de raciocínio, definindo regras mínimas não-redundantes como sendo aquelas que têm antecedente mínimo e conseqüente máximo. Entretanto, Bastide et al. (2000) definem uma restrição adicional: para ser redundante em relação à outra, ela deve ter os mesmos valores de suporte e de confiança da regra não-redundante. Para Aggarwal e Yu (2001), basta que as regras redundantes tenham valor maior ou igual ao das regras não-redundantes. Aggarwal e Yu (2001) também diferenciam as regras redundantes em simples (*simple*), quando a regra redundante tem o mesmo número de itens da regra não-redundante, ou estritas (*strict*), quando ela tem menos itens que a regra não-redundante e o antecedente da regra redundante é um superconjunto do antecedente da regra não-redundante.

Ashrafi et al. (2004) questionam a abordagem de Aggarwal e Yu (2001), argumentando que ela, ao eliminar certas regras, acarreta perda de informação. Os autores adotam uma abordagem mais conservadora, segundo a qual existem apenas dois tipos de regras redundantes: regras com antecedente fixo e conseqüente variável e regras com conseqüente fixo e antecedente variável. Para o primeiro tipo, dada uma regra com mais de um item no conseqüente, se existem as regras com mesmo antecedente, e com cada um dos conseqüentes, então ela não traz informação extra, sendo portanto redundante. O segundo tipo é análogo ao primeiro, só que em relação a um conseqüente fixo e a um antecedente variável.

Ainda seguindo o conceito de min-max (antecedente mínimo, conseqüente máximo), Pasquier et al. (2005) propuseram uma representação concisa de regras de associação usando o conceito de *itemsets* freqüentes fechados. Os autores definiram três maneiras distintas para obter estes conjuntos condensados de regras: min-max exatos, min-max aproximados e min-max aproximados não-transitivos. Para cada abordagem, foi elaborada uma maneira de obter todas as outras regras (externas ao conjunto conciso de regras).

Apesar de interessante, esta abordagem é sensível à ocorrência de *itemsets* freqüentes fechados. Se forem obtidos muitos *itemsets* freqüentes fechados, a capacidade de obter conjuntos condensados de regras é sensivelmente diminuída.

Segundo Zaki (2004), uma regra é não-redundante se ela possui antecedente mínimo e conseqüente mínimo e mesmos valores para suporte e para confiança, diferentemente de Kryszkiewicz (1998), Bastide et al. (2000) e Aggarwal e Yu (2001), que utilizam conseqüente máximo. Desta maneira, se uma regra $X' \rightarrow Y'$ possui os mesmos valores de suporte e de confiança de uma regra $X \rightarrow Y$, e esta última é mais geral que a primeira, isto é, $X \subseteq X'$ e $Y \subseteq Y'$, então $X' \rightarrow Y'$ é dita redundante.

Li e Hamilton (2004) propuseram regras de associação básicas como não-redundantes. Uma regra de associação $X \rightarrow Y$, é dita básica se $\nexists X' \subset X$, tal que, $\forall X''$, onde $X' \subseteq X'' \subseteq X$, $\text{conf}(X'' \rightarrow Y) = \text{conf}(X \rightarrow Y)$. Nesta definição, apenas regras no formato canônico são consideradas: aquelas com um item no conseqüente.

Em muitas situações, a obtenção de *itemsets* freqüentes fechados não diminui satisfatoriamente o número de *itemsets* obtidos pelo fato de a própria técnica ser bastante estrita. Para flexibilizar esta técnica, (Cheng et al., 2006) utilizam um coeficiente delta (δ), de modo a desconsiderar *itemsets* freqüentes cuja relação com o seu menor superconjunto com o maior suporte seja menor que o coeficiente. Os *itemsets* freqüentes fechados restantes são chamados *δ -Tolerance Closed Frequent Itemsets* (δ -TCFI).

Baseados nessa idéia de flexibilizar os *itemsets* freqüentes fechados, Cheng et al. (2008) propuseram obter um conjunto compacto de regras de associação, denominado *δ -Tolerance Association Rules* (δ -TAR). Para isso, definiram o conceito de extensão de suporte, que é um conjunto de valores de suporte estimado para os *itemsets* eliminados pelo coeficiente δ . A partir destes valores de suporte, é possível determinar regras com uma certa precisão nos seus valores de suporte e de confiança. Segundo os autores, a solução reduz bastante o número de regras geradas, em relação a abordagens anteriores, como regras obtidas a partir de *itemsets* freqüentes fechados. Entretanto, essa redução depende do valor de δ : se este for pequeno, ela é menor; se for grande, a redução é maior, porém perde-se precisão dos valores de suporte e de confiança.

Melanda e Rezende (2004) estudaram a capacidade de filtragem de diversas medidas de interesse, e observaram que algumas têm maior capacidade de filtragem. Assim, eles propõem a filtragem com combinações de medidas de interesse, o que acaba por diminuir drasticamente o número de regras geradas, facilitando o entendimento e a utilização do conhecimento pelo especialista do domínio.

Baseado na idéia de obtenção de regras de associação generalizadas apresentada por Srikant e Agrawal (1995), Domingues (2004) propôs generalizar as regras de associação

em uma etapa de pós-processamento, a fim de reduzir o volume de regras geradas. Na sua tese de doutoramento, Carvalho (2007) busca obter conjuntos de regras de associação compactos e representativos usando conhecimento de domínio expresso em taxonomias. Para fazer esta compactação, a abordagem generaliza o antecedente e/ou o conseqüente, desde que todos os itens do atributo generalizado sejam representados no conjunto inicial de regras pelos itens específicos. Apesar de relevante, a solução fica limitada a situações em que há uma taxonomia disponível, o que nem sempre se aplica.

Palshikar et al. (2007) propuseram a redução do número de regras com o conceito de *itemsets* pesados (*heavy itemsets*). Um *itemset* é dito pesado se todas as regras que podem ser obtidas a partir dele estão presentes no conjunto final de regras. Segundo os autores, um conjunto de *itemsets* pesados disjuntos representa um conjunto de regras consideradas não redundantes.

Comparativamente com uma abordagem convencional – aplicando filtros de suporte mínimo e confiança mínima, estas abordagens conseguem produzir conjuntos mais compactos de regras. No entanto, elas têm algumas limitações. As abordagens que se baseiam no conteúdo da regra – (Kryszkiewicz, 1998), (Bastide et al., 2000), (Aggarwal e Yu, 2001), (Ashrafi et al., 2004), (Pasquier et al., 2005), (Zaki, 2004) e (Li e Hamilton, 2004) – têm uma capacidade limitada de redução do número de regras. A baseada em *itemsets* freqüentes fechados (Cheng et al., 2008) considera apenas *itemsets* freqüentes fechados, e não todos, o que faz com que a cobertura de *itemsets* seja baixa. A abordagem baseada em combinações de medidas de interesse (Melanda e Rezende, 2004) utiliza valores rígidos de medidas objetivas para a redução, também desconsiderando a cobertura de *itemsets*. As abordagens baseadas em generalização de regras – (Srikant e Agrawal, 1995), (Domingues, 2004) e (Carvalho, 2007) – dependem da existência de taxonomias. Por fim, a abordagem baseada no conceito de *itemset* pesado (Palshikar et al., 2007) exclui os *itemsets* “quase pesados” (aqueles que geram quase todas as regras possíveis), o que faz com que os conjuntos produzidos não tenham necessariamente uma boa cobertura de *itemsets*. Neste trabalho, propomos um critério de redução de conjuntos de regras que leva em consideração a diversidade de regras, i.e., a cobertura pelas regras dos diferentes *itemsets* induzidos a partir das transações. Esta redução é grande o suficiente para viabilizar a exploração, mas de modo a preservar a cobertura de *itemsets*.

3.3 Abordagens Interativas para a Extração de Regras

Klemettinen et al. (1994) defendem a utilização de *rule templates* para especificar classes de regras interessantes e classes de regras desinteressantes. Os *templates* descrevem con-

juntos de regras especificando quais atributos ocorrem no antecedente e quais ocorrem no conseqüente. Assim, o usuário pode especificar explicitamente o que considera interessante ou não. Para ser interessante, uma regra deve pertencer a uma classe especificada por um *template* de inclusão. Por outro lado, se uma regra pertence a uma classe especificada por um *template* de restrição, ela é considerada não interessante. No mesmo trabalho, os autores propõem técnicas para visualizar as regras, discutidas na **Seção 3.4**.

Goethals e den Bussche (1999) compararam a filtragem *a priori* (por exemplo, utilizando *templates*) com a filtragem *a posteriori*, e concluem que a primeira é mais vantajosa por dois motivos: primeiro, a realização de uma consulta simples é muito mais eficiente, pois o resultado é totalmente direcionado, de modo que apenas regras de interesse são geradas; segundo, se os valores para suporte mínimo e confiança mínima forem baixos, o número de regras geradas será alto, de modo que mesmo a filtragem *a posteriori* pode não ter sucesso.

Goethals e den Bussche (2000) propuseram permitir que o usuário indique restrições, especificadas em uma linguagem de consulta de mineração de dados (*data mining query language*), constituindo sessões de mineração (*mining sessions*). Estas restrições, que podem, por exemplo, determinar os itens do antecedente ou do conseqüente, ou o valor do suporte mínimo ou da confiança mínima, guiam a mineração. Para aplicar estas sessões foram definidas três abordagens: na filtragem *a priori*, cada consulta é executada por uma adaptação do algoritmo que incorpora diretamente as restrições; na filtragem *a posteriori*, o conjunto de regras de associação é minerado e pode, posteriormente, ser consultado; na abordagem híbrida, combina ambas as anteriores.

Baseados na idéia básica de OLAP (*OnLine Analytical Processing*), Aggarwal (2002), Aggarwal e Yu (1998b) propõem a geração *online* de regras de associação. Nessa abordagem, os autores destacaram duas etapas. A primeira é uma etapa de pré-processamento, em que os dados são armazenados em um *lattice* de adjacência com informações sobre *itemsets* freqüentes, de acordo com um determinado limiar (número máximo de *itemsets* freqüentes). A partir deste *lattice* de adjacência, é possível obter respostas instantâneas a consultas *online*. A segunda etapa consiste na eliminação de regras de associação redundantes, abordado na **Seção 3.2**.

Hidber (1999) apresenta um algoritmo *online* para computar *itemsets* freqüentes, em que o usuário pode ajustar livremente o nível de suporte durante a mineração. Armazenando um superconjunto dos *itemsets* freqüentes e um intervalo determinístico do suporte para cada *itemset*, o algoritmo produz um resultado correto após duas varreduras da base de transações.

Jorge (2004) propôs um método para agrupar e sumarizar grandes conjuntos de regras de associação de acordo com seu conteúdo, aplicando agrupamento hierárquico para particionar o conjunto inicial de regras em conjuntos temáticos. Para obter os agrupamentos, é definida uma distância métrica binária entre as regras. A partir daí, o conjunto de regras é sumarizado, escolhendo-se uma regra de cada grupo como representativa a fim de facilitar a tarefa de exploração interativa pelo usuário. Esta abordagem, no entanto, prevê o agrupamento de regras, e não de *itemsets*, de maneira que o processo não pode ser direcionado durante a sua execução.

As abordagens apresentadas possibilitam maneiras diferentes de interação do minerador com o processo, fornecendo subsídios para que ele tenha um papel mais ativo, podendo explorar com mais facilidade conjuntos de regras de associação. Entretanto, elas apresentam limitações em cenários exploratórios, em que o minerador não tem pleno conhecimento das bases de dados. Nas abordagens baseadas em *templates* (Klemettinen et al., 1994), (Goethals e den Bussche, 2000) ou na idéia básica de *OLAP* (Aggarwal, 2002), (Aggarwal e Yu, 1998b), o minerador precisa ter diretrizes claras para iniciar a mineração. A abordagem *online* (Hidber, 1999) permite ajustar o nível de suporte, porém não permite que o minerador infira correlações durante o processo, limitando a sua capacidade de análise. Por fim, a abordagem de agrupamento e sumarização de regras (Jorge, 2004) pode ser um bom meio de organização de conjuntos de regras, porém não permite que o minerador direcione o processo durante a sua execução.

3.4 Abordagens Interativas com Visualização para a Exploração de Regras

Representações visuais típicas para exibir resultados da tarefa de obtenção de regras de associação adotam estruturas do tipo matriz, gráficos de barras e combinações de ambas. Nas primeiras, geralmente bidimensionais, cada célula corresponde a uma regra, com uma das dimensões representando antecedentes das regras, e a outra representando os conseqüentes. Por sua vez, os gráficos de barras são usados para representar valores de medidas de interesse, por exemplo, suporte e confiança, de maneira que a altura da barra seja diretamente proporcional ao seu valor.

Na **Figura 3.1** é mostrado um exemplo de matriz bidimensional. Na **Figura 3.1(a)**, é mostrada a regra $B \rightarrow C$, com os valores de suporte e de confiança mostrados como as barras azul e vermelha. Esta representação é bastante simples e intuitiva, porém apresenta algumas limitações. Uma delas diz respeito à dificuldade de se representar antecedentes e conseqüentes com mais de um item. Uma maneira de fazer isso é utilizando-se mais de uma

célula para representar uma regra, como mostrado na **Figura 3.1(b)**, em que a regra $A, B \rightarrow C$ é representada. Porém, esta representação é ambígua, pois pode representar também duas regras: $A \rightarrow C$ e $B \rightarrow C$. Outro problema enfrentado por este tipo de representação é a possibilidade de haver a oclusão de objetos, que pode dificultar a visualização dos mesmos, como mostrado na **Figura 3.1(c)**.

Figura 3.1: Visualização básica de regras usando matriz.

Com o objetivo de tratar estes problemas, Wong et al. (1999) introduziram uma representação que mapeia relacionamentos item-regra em vez de item-item (neste caso, antecedente-conseqüente). Nesta representação, os itens são representados em linhas e as regras, em colunas. Cada relacionamento item-regra é representado por um bloco posicionado na interseção entre a linha (item) e a coluna (regra). O bloco pode ser azul ou vermelho: o primeiro indica que o item faz parte do antecedente da regra, enquanto que o segundo indica que o item faz parte do conseqüente. Desta maneira, elimina-se o problema da ambigüidade, pois é possível representar quaisquer combinações entre itens. Além disso, os valores de suporte e de confiança são representados em gráficos de barra ao fundo da matriz, em uma visão tridimensional, eliminando o problema de oclusão de objetos. A **Figura 3.2** mostra um exemplo de aplicação da abordagem a uma base de dados de notícias.

Apesar de eliminar diversas limitações, a abordagem de Wong et al. (1999) não favorece a identificação das regras quando o número delas é muito grande. Para identificar uma regra, percorre-se uma determinada coluna, de modo a identificar todos os blocos azuis ou vermelhos. No caso de centenas ou de milhares de itens, é necessário construir uma matriz com centenas ou milhares de itens, o que extrapola o espaço de visualização.

Chakravarthy e Zhang (2003) também propuseram uma matriz bidimensional para visualização de regras de associação. A idéia é dividir o processo em dois passos. No primeiro, as regras são agrupadas segundo o número de itens do antecedente (uma barra

Figura 3.3: Visualização de regras, por Chakravarthy e Zhang (2003).

Techapichetvanich e Datta (2005) propuseram a visualização de regras de associação utilizando uma estrutura que lembra uma matriz. Nesta abordagem, chamada VisAR, que também utiliza *templates*, a idéia é mostrar os itens do antecedente e do conseqüente em uma dimensão, separados por uma linha, e as regras, em uma outra dimensão. Desta maneira, as relações entre itens e regras são determinadas por nós que são desenhados nas interseções de maneira apropriada. Nesta abordagem, os valores para o suporte ou para a confiança são divididos em intervalos e mapeados pela cor da linha da regra. Esta abordagem é ilustrada na **Figura 3.4**. No exemplo, dez regras cujo antecedente é *cd, rice* são representadas. *newspaper*, *soya sauce*, *battery* e *sweets* são itens pertencentes ao conseqüente de pelo menos uma das regras. O suporte (alternativamente, a confiança) é mapeado utilizando a escala de cores para a linha vertical correspondente à regra.

Apesar de aparentemente utilizar menos espaço de tela, por conta da diminuição do espaço alocado para cada coluna, esta abordagem também sofre severamente da utilização de espaço: itens que estejam no antecedente de pelo menos uma regra e no conseqüente de pelo menos uma regra (que não seja a mesma) são listados duas vezes. Desta maneira, para n itens existentes no conjunto de regras, pode-se ter até $2*n$ linhas para representá-los.

A abordagem de Klemettinen et al. (1994) utiliza gráficos de barra. No entanto, ao invés de uma matriz bidimensional, é exibida uma lista de regras, sendo que três barras são apresentadas para cada regra, representando as medidas de suporte, confiança e o produto

Figura 3.4: Visualização de regras, por Techapichetvanich e Datta (2005).

entre elas, chamado *commonness*. Esta representação é mostrada na **Figura 3.5(a)**. Para reduzir o tamanho da lista, os autores utilizaram um mecanismo de *templates*.

Neste trabalho, os autores também apresentaram uma visualização de regras baseada em grafos, na qual os itens são representados por nós e as associações, por arcos dirigidos (do antecedente e para o conseqüente), como mostrado na **Figura 3.5(b)**. É possível representar regras com dois ou mais itens no antecedente (por exemplo, regra C,D $\rightarrow$ B). Entretanto, regras com dois ou mais itens no conseqüente não são previstas. A espessura do arco representa a confiança ou o suporte da regra. As cores podem fornecer informação adicional, como o valor do *commonness*.

Figura 3.5: Visualização de regras, por Klemettinen et al. (1994).

Apesar de utilizar mais eficientemente o espaço de desenho comparado às abordagens baseadas em matriz, há um problema de densidade do grafo quando muitas regras são representadas. Para minimizar este problema é utilizado um mecanismo de *templates*

para reduzir o número de regras. Outra possível solução é limitar o tamanho das regras, por exemplo, em termos do número de itens no antecedente. Outras idéias incluem juntar nós em um agrupamento, transformando-o em um nó único, ou ainda separar os nós que representam antecedentes dos que representam conseqüentes, colocando-os em regiões separadas da área de desenho. A abordagem também é limitada por não representar adequadamente conseqüentes com dois ou mais itens.

Hofmann et al. (2000) propuseram uma representação usando *Mosaic Plots* e *Double Decker Plots*, que mostram uma tabela de contingência representando uma regra. Ao mapear o suporte e a confiança de cada regra a áreas proporcionais aos seus valores, a representação provê um entendimento mais profundo da correlação entre o antecedente e o conseqüente (e suas negativas) da regra.

É representada na **Figura 3.6** a regra *corned beef, avocado, artichoke* $\rightarrow$ *heineken*. Além da regra propriamente dita, a representação apresenta todas as combinações de ocorrência ou não ocorrência de cada um dos itens envolvidos, representadas pelas barras cheias (preto) ou vazias (branco) na base da figura. No exemplo, quatro combinações se destacam pelo alto valor de confiança. Uma delas, $\neg$ *corned beef, avocado, artichoke* $\rightarrow$ *heineken* se destaca por ter um alto valor de suporte e o maior valor de confiança (quarta coluna, da esquerda para a direita).

Figura 3.6: Visualização de regras, por Hofmann et al. (2000).

Apesar de auxiliar na análise de uma regra individual, esta abordagem não proporciona uma visão geral de um conjunto de regras. Além disso, a interpretação da visualização pode se tornar não-trivial quando há regras com muitos itens no antecedente e/ou no conseqüente.

Uma visualização por gráfico de dispersão do corpus de regras de associação chamada *TwoKey plot*, foi proposta por Unwin et al. (2001). Nesta visualização, mostrada na

Figura 3.7, as duas medidas de interesse mais comuns (suporte e confiança) são mapeadas para os dois eixos. Assim, a posição de uma regra no gráfico é determinada pelos valores de suporte e de confiança, enquanto que a cor mapeia o número de itens da regra. Uma vez que o minerador tem uma idéia geral do conjunto de regras como um todo, ele/a pode direcionar a exploração ajustando os valores de suporte mínimo e de confiança mínima, limitando a área de visualização. A solução inclui mecanismos de consulta para obter regras relacionadas, como filhos/descendentes (dada uma regra, recuperar regras mais específicas de acordo com o antecedente), pais/antecedentes (dada uma regra, recuperar regras mais gerais de acordo com o antecedente), e vizinhos (dada uma regra, recuperar regras a uma certa distância d).

Figura 3.7: Visualização de regras, por Unwin et al. (2001).

Diferentemente das abordagens que utilizam como parâmetros de representação os itens do antecedente e do conseqüente, esta representação utiliza os valores de suporte e de confiança, o que a torna adequada para a utilização de mecanismos de filtragem. Entretanto, este tipo de abordagem não é favorável se o minerador necessita focar no conteúdo. Além disso, é preciso tratar os casos de regras que têm os mesmos valores das medidas de interesse, e são mapeadas na mesma posição.

Representações visuais de regras de associação ainda incluem abordagens baseadas em grades, como em (Ong et al., 2002). Neste tipo de representação, colunas são dispostas de acordo com valores crescentes de confiança, e linhas dispostas em ordem crescente de suporte, como mostrado na **Figura 3.8(a)**. As regras são posicionadas e agrupadas em células, de acordo com os valores de suporte e de confiança, e representadas por quadrados. O suporte, a confiança e filtros baseados nos itens ajudam o minerador a navegar pelo conjunto de regras. Os autores apontam que a abordagem é adequada para regras de qualquer tamanho. Entretanto, apenas os valores das medidas de interesse são visualmente representados.

Outra representação proposta por Ong et al. (2002) para visualizar regras de associação é uma estrutura hierárquica (ver **Figura 3.8(b)**), na qual as regras são agrupadas de acordo com o antecedente e o conseqüente. Primeiramente, os antecedentes são mostrados como nós, segundo o número de itens. Para cada nó (antecedente), as regras contendo os conseqüentes são mostrados como subnós. As duas representações estão presentes em um protótipo chamado *CrystalClear*.

Figura 3.8: Protótipo CrystalClear, por Ong et al. (2002).

Assim como a abordagem de Unwin et al. (2001), esta utiliza como parâmetro de representação os valores de medidas de interesse. Deste modo, esta abordagem também precisa agrupar as regras, fazendo-o em quadrados. A representação hierárquica, por sua vez, favorece a inspeção de regras tomando como ponto de partida o antecedente. No entanto, a inspeção pelo conseqüente necessita de uma segunda representação.

Kopanakis e Theodoulidis (2003) apresentaram uma representação distinta para regras de associação usando gráficos de barra, em conjunto com coordenadas paralelas (Inselberg e Dimsdale, 1990) e uma representação baseada em grade. A abordagem para os gráficos de barra consiste em mapear não só os valores das medidas de interesse, mas também o antecedente e o conseqüente das regras. Cada elemento, discreto ou contínuo, tanto do antecedente quanto do conseqüente, é mapeado visualmente como uma barra. Se o elemento é discreto, então o valor é indicado na barra; se é contínuo, um intervalo é indicado. Na **Figura 3.9**, é mostrada a regra $10 \leq A \leq 25, B \text{ IN } \{B1\}, 40 \leq C, D \text{ IN } \{D1, D2\} \rightarrow 0 \leq E \leq 45, 0 \leq CL \leq 50$, em que os componentes A, C, E e CL são contínuos, enquanto que B e D são discretos.

Esta abordagem para representar uma única regra consegue mapear atributos discretos, contínuos e intervalares. Entretanto, intervalos não são usualmente utilizados para o

Figura 3.9: Gráfico de barras, por Kopanakis e Theodoulidis (2003).

problema de obtenção de regras de associação, mas para classificação. Para representar várias regras, uma representação baseada em grade é empregada, em que cada regra é mapeada em uma célula diferente, aplicando-se uma medida de similaridade entre as regras para determinar o seu posicionamento, de modo que regras com conteúdo similar estejam próximas umas das outras.

A segunda representação utiliza a técnica de coordenadas paralelas. Os itens são mapeados em eixos paralelos, e cada regra é representada como uma polilinha intersectando cada eixo em um ponto correspondendo ao seu valor. Na **Figura 3.10** é representada a regra *Account_Status* IN {*Balanced*, *DaysLate30*}, *Marital_Status* IN {*Single*}, $466,00 < Mo_Expenses \rightarrow Home = Own$, *Credit_Limit* = *high*, em que os atributos *Account_Status*, *Marital_Status*, *Home* e *Credit_Limit* são discretos, e *Mo_Expenses* é contínuo.

Figura 3.10: Coordenadas paralelas, por Kopanakis e Theodoulidis (2003).

A terceira representação, em grade, oferece uma perspectiva diferente para visualizar duas ou mais regras simultaneamente, pelo mapeamento de um atributo para cada célula

da grade, e uma regra para cada linha. Assim, uma célula corresponde a uma sub-expressão (valor discreto, contínuo ou intervalar), com a existência de valores dentro da célula indicando a sua presença na regra. O suporte e a confiança são mostrados no lado esquerdo. Na **Figura 3.11**, são apresentadas dezesseis regras envolvendo treze atributos.

Figura 3.11: Gráfico baseado em grade, por Kopanakis e Theodoulidis (2003).

Dentre as três representações propostas, esta aparenta ser a mais completa, já que consegue mapear todos os tipos de atributos, além de mostrar várias regras ao mesmo tempo, permitindo compará-las. É possível perceber na figura anterior dois grandes grupos: o primeiro, contendo as regras R_1 a R_10; o segundo, as regras R_11 a R_16.

Na representação visual baseada em grafos por Bruzzese e Buono (2004), itens são representados por nós, sendo que a sua cor indica se é antecedente (em vermelho) ou o conseqüente (em azul) de uma regra, enquanto que as regras de associação são representadas por arestas. A confiança da regra é mapeada pelo comprimento da aresta, enquanto que a sua cor mapeia o suporte – azul claro e azul escuro se referem a valores baixos e altos de suporte, respectivamente. Um exemplo é mostrado na **Figura 3.12(a)**.

O minerador pode explorar mais a fundo subconjuntos de regras por meio de uma representação por coordenadas paralelas, cujo objetivo é representar regras de associação em termos da utilidade de cada item do antecedente da regra, dada pela expressão $IU_i = \frac{C_R - C_{R(\neg y)}}{\max(C_R; C_{R(\neg y)})}$. Segundo os autores, a relevância de um item pode ser medida pela sua utilidade. Desta maneira, dada a seleção de um conseqüente, é apresentada uma visualização em que cada eixo representa um item do antecedente. Para cada regra, uma polilinha conectando a utilidade de cada item é desenhada. Desta maneira, é possível identificar visualmente os itens com menores valores de relevância, descartando-os,

enquanto seleciona-se aqueles com os maiores valores de relevância, como ilustrado na **Figura 3.12(b)**.

(a) Gráfico baseado em grafo

(b) Coordenadas paralelas

Figura 3.12: Visualização de regras, por Bruzzese e Buono (2004).

Apesar de ser menos sensível ao problema de uso de espaço de desenho, a representação baseada em grafos de Bruzzese e Buono (2004) apresenta uma limitação séria quanto a este aspecto. Nesta representação, cada nó representa um conjunto de itens (e não um item). Desta maneira, são necessários muitos nós para representar todas as combinações possíveis de itens, tanto para o antecedente, quanto para o conseqüente.

A representação por coordenadas paralelas também tem uma limitação, herdada da própria técnica: quando a quantidade de polilinhas é muito grande, a identificação dos objetos é severamente prejudicada devido à enorme sobreposição de linhas.

Kuntz et al. (2000) introduziram uma nova abordagem, centrada no usuário, para extração de regras de associação. Partindo de itens freqüentes, como no algoritmo *Apriori*, ao selecionar um item de interesse, o usuário pode obter regras envolvendo o item selecionado e outros itens relacionados. Esta operação é chamada *forward chaining* (*backward chaining*, a operação inversa, também é possível). Os autores também empregam uma visualização baseada em grafos para apresentar visualmente estas operações. Esta estratégia para extrair regras, inspirada em uma navegação, é dirigida pelo usuário, e evita a extração de regras não interessantes. Um exemplo é mostrado na **Figura 3.13**. O minerador partiu do item *o11*, representado por um nó à esquerda, e executou a operação *forward chaining*, a fim de obter possíveis combinações deste item com outros, obtendo quatro *itemsets* como resposta: $\{o11, h11\}$, $\{o01, o11\}$, $\{o11, h01\}$ e $\{o11, o15\}$, mostrados no grafo como nós conectados a *o11*. Em um segundo momento, o minerador

decidiu prosseguir na exploração de $\{o01, o11\}$, procedendo da mesma maneira e obtendo quatro outros *itemsets*: $\{o01, o11, h01\}$, $\{o01, o11, h11\}$, $\{o01, o02, o11\}$ e $\{o01, o11, o15\}$. Continuando com a exploração sobre este último *itemset* ($\{o01, o11, o15\}$), ele obteve: $\{o01, o11, o15, h01\}$, $\{o01, o11, o15, h15\}$, $\{o01, o11, o15, h11\}$ e $\{o01, o02, o11, o15\}$.

Figura 3.13: Visualização de regras, por Kuntz et al. (2000).

Esta abordagem favorece a exploração não-linear dos itens, combinando-os de acordo com a necessidade do minerador. Ela direciona a obtenção de regras, reduzindo o número de regras extraídas. Entretanto, esta abordagem, assim como o uso de consultas ou de *templates*, requer que o minerador tenha previamente uma direção a seguir na exploração. Em contextos em que isto não ocorre, por exemplo, quando o domínio é desconhecido, a abordagem não é adequada.

Blanchard et al. (2003b,a) propuseram uma exploração de regras centrada no usuário, que adota uma metáfora visual de duas metades de tigela para posicionar as regras de interesse (neste caso, regras selecionadas pelo minerador), ilustrado na **Figura 3.14**. Enquanto uma das metades contém as regras generalizadas, a outra contém as regras especializadas. Cada regra é representada por uma esfera, cujo raio mapeia o valor do seu suporte, e por um cone, cujo diâmetro da base mapeia a sua confiança. Além disso, a cor da esfera e do cone representam redundantemente uma média ponderada dos valores destas medidas. A posição da regra em uma das metades da tigela da metáfora representa a intensidade da implicação.

Este trabalho é estendido em (Blanchard et al., 2007b) com a adição de duas visualizações: a primeira é uma metáfora visual 3D e a outra é uma visualização 2D de

Figura 3.14: Visualização de regras, por Blanchard et al. (2003b).

conjuntos de regras e suas relações de vizinhança (Kuntz et al., 2000). Baseado nas relações de vizinhança, os autores definiram regras que são mais próximas de uma regra selecionada, de acordo com critérios predefinidos. Estas relações são: *same antecedent*, *forward chaining*, *antecedent generalization* (o oposto de *forward chaining*) e *same items*.

Em outra extensão deste trabalho (Blanchard et al., 2007a), foi criada a interface ARVis, na qual restrições sintáticas e de medidas de interesse podem ser aplicadas às regras. As restrições sintáticas especificam as restrições de ocorrência de itens no antecedente e/ou no conseqüente, enquanto que as de medidas de interesse especificam intervalos válidos para as medidas de interesse. As relações de vizinhança de (Blanchard et al., 2007b) também foram adicionadas à interface, juntamente com outras que foram criadas: *agreement specialization*, *exception specialization*, *generalization* e *same consequent*.

Couturier et al. (2007a) combinam diversas representações visuais de regras de associação com a técnica de distorção *fish-eye view*, para permitir que o minerador tenha acesso visual tanto ao contexto geral quanto a detalhes sob demanda. Em uma das visualizações, as regras são exibidas em uma matriz 2D, tendo antecedentes mapeados em um eixo e conseqüentes mapeados em outro, e valores para as medidas de interesse (suporte e confiança) mapeados em cores, dentro das respectivas células.

Os autores propõem em (Couturier et al., 2007b) a exploração visual de regras de associação por meio de representações visuais de agrupamentos de regras. Eles empregaram o algoritmo *k*-médias sobre as características das regras para determinar os agrupamentos. Cada agrupamento é representado em uma célula da grade, e cores representam valores para as medidas de interesse, como mostrado na **Figura 3.15**.

Figura 3.15: Visualização de regras, por Couturier et al. (2007a,b).

Poucas abordagens provêm uma representação visual de uma regra por vez, como as de Hofmann et al. (2000) e Kopanakis e Theodoulidis (2003). Elas têm a vantagem óbvia de usar espaços de tela reduzidos, mas não apoiam a comparação entre múltiplas regras.

A maior parte das abordagens que criam representações visuais de múltiplas regras não consegue evitar o ponto de saturação visual, a partir do qual regras não podem ser mais representadas na tela ou a compreensão da representação é seriamente prejudicada. Isto ocorre nas representações baseadas em matrizes bidimensionais (Wong et al., 1999), (Chakravarthy e Zhang, 2003), (Couturier et al., 2007a), (Couturier et al., 2007b) ou em grades (Techapichetvanich e Datta, 2005), (Kopanakis e Theodoulidis, 2003).

Por outro lado, as representações visuais baseadas em grafos (Klemettinen et al., 1994), (Kuntz et al., 2000), (Bruzese e Buono, 2004), assim como as baseadas em projeções (Unwin et al., 2001), (Ong et al., 2002) são menos sensíveis ao problema de saturação, mas podem sofrer de oclusão de objetos na tela. Em outras palavras, existe uma sobreposição dos marcadores gráficos usados para representar regras, o que reforça a necessidade de prover mecanismos de interação para contornar o problema.

A representação proposta por Blanchard et al. (2007a,b) também sofre do uso de muito espaço na tela – uma alternativa para contornar este problema é mudar a posição da câmera, porém esta política pode diminuir o tamanho dos objetos, prejudicando a percepção.

Uma possibilidade para evitar estes problemas é fornecer uma visão geral do conjunto de regras, que não deve ser muito grande, que permita selecionar subconjuntos de regras de interesse que podem ser explorados por um método que provêm informações detalhadas.

Esta abordagem vai ao encontro do mantra da busca visual de informação, descrito por Shneiderman (1996): “*Overview first, zoom and filter, then details-on-demand*”.

3.5 Abordagens para Visualização durante a Obtenção de Regras

As abordagens da seção anterior, voltadas para a visualização de resultados de tarefas de obtenção de regras de associação, apoiam a busca por regras interessantes dentro de enormes conjuntos de regras utilizando representações visuais interativas. Entretanto, os conjuntos de regras já previamente obtidos não possibilitam a atuação do minerador para direcionar a execução do algoritmo. Conhecimento prévio sobre o domínio permite direcionar os conjuntos de regras extraídos para um determinado objetivo e reduzir o número de regras. Além disso, o minerador pode adquirir mais conhecimento do domínio tendo uma interação mais intensa com o processo. Nesta seção, são apresentadas abordagens que adotam metáforas visuais para apoiar a intervenção do usuário durante o processo de obtenção de regras de associação.

Uma interface visual para mostrar o *lattice* obtido durante a extração de regras foi sugerida por Mahanti e Alhajj (2005). Esta interface, chamada *Visual Apriori* (**Figura 3.16**), mostra a criação e a eliminação de nós da estrutura, com foco voltado para o ensino do algoritmo. *Itemsets* freqüentes são apresentados, porém as regras de associação em si não são. Trata-se de uma representação valiosa porque permite ao minerador monitorar a construção do *lattice*, o que é importante para construir um modelo mental da execução do algoritmo, ao mesmo tempo em que apresenta uma visão geral do processo de obtenção de *itemsets* freqüentes. Apesar destes pontos positivos, o *Visual Apriori* não permite que o minerador tenha um papel ativo no processo, por exemplo, não permite a eliminação de nós ou a mudança de parâmetros de execução.

Zaki e Phoophakdee (2003) propuseram a visualização de *itemsets* freqüentes fechados em um sistema, denominado MIRAGE (*Interactive Graphical Exploration of Minimal Association Rules*), que emprega uma visualização compacta baseada em grafos de *itemsets* freqüentes fechados, como mostrado na **Figura 3.17**. A partir daí, permite-se gerar regras de associação. A visualização é separada em camadas, de acordo com o número de itens de cada *itemset* freqüente fechado. Cada *itemset* freqüente fechado é mapeado como um nó e uma linha conecta dois *itemsets* freqüentes fechados se um deles é subconjunto de outro e não existe *itemset* freqüente fechado intermediário entre eles e que possa ser conectado a um deles.

Figura 3.16: Visualização de *itemsets* frequentes, por Mahanti e Alhajj (2005).

Figura 3.17: Visualização de *itemsets* frequentes, por Zaki e Phoophakdee (2003).

Yang (2003) sugeriu a visualização de *itemsets* frequentes e de regras de associação usando coordenadas paralelas. Primeiramente, os itens são separados em grupos para que sejam dispostos de maneira consecutiva na representação visual, segundo a frequência. Daí, n eixos paralelos são desenhados, tal que n é o número de itens do maior *itemset*. Os *itemsets* frequentes são desenhados como polilinhas conectando os eixos nas posições correspondentes aos itens que possui. As regras são representadas de maneira similar: duas polilinhas, uma para o antecedente, e outra para o conseqüente são desenhadas. Uma seta conecta as polilinhas, e a cor pode mapear os valores de medidas de interesse. Um exemplo é mostrado na Figura 3.18.

Figura 3.18: Visualização de *itemsets* frequentes e de regras, por Yang (2003).

Ertek e Demiriz (2006) empregaram uma técnica baseada em grafos para mapear *itemsets* frequentes e regras. Nesta representação, itens e *itemsets* são mapeados como nós e seus interrelacionamentos são mapeados como conexões entre esses nós (**Figura 3.19(a)**). Para visualizar regras, itens e *itemsets* também são mapeados como nós, e conexões entre itens e regras mapeiam as relações entre eles. Estas conexões são dirigidas para indicar a direção da implicação (**Figura 3.19(b)**).

Figura 3.19: Visualização de *itemsets* frequentes e de regras, por Ertek e Demiriz (2006).

A abordagem por Zaki e Phoophakdee (2003) apenas gera os resultados – os *itemsets* freqüentes fechados – e os exhibe. As abordagens de Yang (2003) e de Ertek e Demiriz (2006) geram os resultados e exibem tanto os resultados intermediários – os *itemsets* freqüentes – como os resultados finais – as regras de associação. Entretanto, eles não permitem que o usuário interfira no processo. O trabalho de Liu e Salvendy (2006) dá um passo nesta direção, permitindo a exploração de *itemsets* freqüentes ao final de cada passo do algoritmo *Apriori* (**Figura 3.20**). Eles empregam uma estrutura visual baseada em uma matriz bidimensional para representar *itemsets* freqüentes, onde itens e *itemsets* são mapeados em linhas e colunas, respectivamente, de modo que cada relação item-*itemset* é mapeado na posição que os intersecta. Eles também empregam uma visualização de regras, similar à visualização de *itemsets*: itens e regras são mapeados em linhas e colunas, mas um “A” ou um “C” na posição de interseção indica se o item pertence ao antecedente ou ao conseqüente da regra.

Figura 3.20: Visualização de *itemsets* freqüentes e de regras, por Liu e Salvendy (2006).

A abordagem proposta por Liu e Salvendy (2006) permite a geração interativa de *itemsets* freqüentes e a geração de regras a cada passo do *Apriori*, mas o minerador não tem total controle sobre o processo por duas razões. Primeiro, porque as regras são geradas automaticamente a partir de todos os *itemsets* freqüentes obtidos até então. Deste modo, o minerador não pode focar na exploração de grupos específicos de *itemsets*. Além disso, os mineradores não têm controle sobre a execução do algoritmo, visto que o processo só pode avançar, não pode retroceder. Em se tratando de abordagens interativas

para explorar *itemsets* ou mesmo regras de associação, é importante que os mineradores possam desfazer ações, de modo a tentar caminhos alternativos de execução, buscando obter resultados mais satisfatórios. Por exemplo, se o minerador ajusta alguns parâmetros e mais adiante no processo os julga inadequados, é útil voltar algumas ações (desfazê-las) e retomar o processo usando outros parâmetros. Uma maneira de lidar com este problema é permitir que o minerador recomece (reinicie o processo), desfça ações (volte), avance (em direção ao fim do processo), ou execute completamente (até o fim). Na **Tabela 3.1**, é mostrado um resumo das principais características das abordagens discutidas.

Tabela 3.1: Tabela comparativa das abordagens de visualização de *itemsets*.

	Mahanti & Alhajj (2005)	Yang (2003)	Liu & Salvendy (2006)	Ertek & Demiriz (2006)
Representação de <i>itemsets</i> com 1 item	Boa (primeiro nível da estrutura <i>trie</i>)	(não definido)	Boa (uma marca na interseção da linha com a coluna)	Boa (uma conexão entre um nó-item e um nó- <i>itemset</i>)
Representação de <i>itemsets</i> com 2+ itens	Ruim (usuário precisa percorrer a estrutura)	Boa (polilinhas longas indicam <i>itemsets</i> com mais itens)	Regular (cada relacionamento item- <i>itemset</i> é representado por uma marca)	Boa (nó adicional e arcos ligando-o aos itens; cor mapeia tamanho)
Representação do suporte de <i>itemsets</i>	Regular (valor é mostrado ao lado do <i>itemset</i>)	Boa (valor é mostrado na seleção; cor é indicada para cada segmento)	Boa (barras e valores são mostrados, na seleção)	Boa (tamanho do nó- <i>itemset</i> indica o valor)
Representação do tamanho de <i>itemsets</i>	Regular (indicado pelo número de nós percorridos)	Boa (indicado pelo número de eixos tocados pela polilinha)	Regular (indicado pelo número de marcas na coluna correspondente)	Boa (indicado pela cor do nó- <i>itemset</i> correspondente)
Escalabilidade visual	Ruim (toda a estrutura <i>trie</i> é representada)	Ruim (para cada <i>itemset</i> , uma polilinha é adicionada)	Ruim (o tamanho da matriz é dado por n <i>itemsets</i> vezes m itens)	Ruim (para cada <i>itemset</i> , um nó e arestas ligando-o aos seus itens são adicionados)
Seleção e exploração de um <i>itemset</i>	Ruim (o usuário precisa percorrer a estrutura)	Ruim (clicando na polilinha correspondente)	Ruim (o usuário precisa varrer toda a matriz e identificar os itens)	Boa (clicando no nó- <i>itemset</i> , as conexões indicam os seus itens)
Seleção, exploração e comparação de dois ou mais <i>itemsets</i>	Ruim (o usuário precisa percorrer a estrutura)	Regular (a proximidade e o ângulo dos segmentos das polilinhas indica similaridade)	Ruim (o usuário precisa varrer toda a matriz e comparar item por item)	Ruim (o usuário precisa selecionar um <i>itemset</i> por vez e comparar seus itens)

3.6 Considerações Finais

Neste capítulo, foi feito um levantamento de abordagens que buscam tratar os problemas levantados no **Capítulo 1** adotando as medidas a seguir: reduzindo o tamanho dos conjuntos de regras, flexibilizando o processo, aumentando a interação do usuário, ou provendo representações visuais para auxiliar na exploração de conjuntos de regras ou de *itemsets* freqüentes.

Entretanto, estas abordagens não são adequadas para algumas situações – por exemplo, quando o usuário não conhece muito o domínio dos dados (cenário exploratório) ou quando ele/a tem conhecimento sobre o domínio, mas quer obter conhecimento além do esperado. As abordagens que reduzem o tamanho dos conjuntos de regras não se preocupam em reduzir a redundância das regras, especialmente em cenários exploratórios. As que flexibilizam o processo são passivas, no sentido de que apenas provêm conhecimento especificado pelo usuário, não instigando-o com conhecimento adicional possivelmente útil. As abordagens que provêm representações visuais de regras não podem ser utilizadas durante o processo, o que é essencial para o monitoramento e direcionamento do mesmo. Por fim, as que provêm representações visuais de *itemsets* freqüentes são limitadas, pois não provêm informação clara sobre agrupamentos de *itemsets* freqüentes, tampouco permitem ao usuário navegar pelas etapas nos dois sentidos (avanço e retrocesso) ou ainda mudar os parâmetros de execução.

No próximo capítulo, é apresentada uma abordagem para mineração visual de dados que considera a diversidade de regras para filtragem de conjuntos de regras de associação, além de permitir que o usuário direcione o processo e permitir a exploração visual de regras de associação por comparação por pares. É também apresentado no próximo capítulo o **Sistema I₂E**, um ambiente computacional projetado segundo esta abordagem.

Abordagem Interativa para Extração e Exploração de Regras

4.1 Considerações Iniciais

Neste capítulo, é introduzida uma abordagem para a obtenção de regras de associação que parte da possibilidade do usuário de direcionar o processo, chamada **Abordagem I₂E**. A abordagem, apoiada pelo **Sistema I₂E** apresentado posteriormente neste capítulo, é dividida em duas grandes etapas, correspondendo às etapas de extração e de exploração de regras.

Para a primeira etapa, é apresentada uma metodologia para a extração interativa de regras, na qual o minerador pode executar interativamente o algoritmo *Apriori*, visualizar os *itemsets* freqüentes à medida que eles são gerados ao final de cada passo, e obter conjuntos compactos de regras, tendo como base para a geração o *itemset* gerador, e como base para a filtragem, grupos de regras provenientes de *itemsets* similares. Esta metodologia, apoiada pelo **módulo I₂E** do sistema citado, permite redirecionar o processo, seja mudando os *itemsets* de interesse, eliminando os desinteressantes, seja mudando dinamicamente o valor de suporte mínimo ao final de cada etapa do *Apriori*.

Para a segunda etapa, apoiada pelo **módulo RuleX**, é apresentada uma estratégia que direciona a exploração de regras, na qual o minerador pode avaliar seu interesse pelas regras por meio de comparações entre regras, com o auxílio de representações visuais.

Inicialmente alguns termos são definidos. Em seguida, é relatado um estudo preliminar em que foram estudados os resultados de uma mineração convencional, e que direcionou a solução proposta nesta tese. Depois, as duas etapas da abordagem são descritas e discutidas. Por fim, é apresentado um modelo interativo, centrado no usuário, para tarefas de obtenção de regras de associação.

Definição 4.1 *Seja I um itemset freqüente e r uma regra de associação gerada a partir de I , então I é o **itemset gerador** de r .*

Definição 4.2 *Seja I um itemset freqüente e seja $C = \{c_1, c_2, \dots, c_i\}$ o conjunto de i itens nele contidos. Diz-se que o **tamanho do itemset** I é dado por $|C|$, isto é, o número de elementos pertencentes ao conjunto C .*

Definição 4.3 *Seja r uma regra de associação e seja $C' = \{c_1, c_2, \dots, c_i\}$ o conjunto de i itens nela contidos. O **tamanho da regra de associação** r é dado por $|C'|$, isto é, o tamanho do conjunto C' .*

Definição 4.4 *Sejam I_1 e I_2 dois itemsets, a **interseção** entre eles é dada por $I_1 \cap I_2 = \{i | i \in I_1 \text{ e } i \in I_2\}$.*

Definição 4.5 *Sejam I_1 e I_2 dois itemsets, a **união** entre eles é dada por $I_1 \cup I_2 = \{i | i \in I_1 \text{ ou } i \in I_2\}$.*

Definição 4.6 *Seja uma regra r do tipo $A \rightarrow B$. Indica-se por **antec(r)** o antecedente da regra r , que é dado por A .*

Definição 4.7 *Seja uma regra r do tipo $A \rightarrow B$. Indica-se por **conseq(r)** o conseqüente da regra r , que é dado por B .*

4.2 Identificação do Problema – Extração de Regras Representativas

Um estudo preliminar foi conduzido com o objetivo de analisar as características de conjuntos de regras de associação obtidos por um processo convencional, não interativo de obtenção de regras de associação. Este estudo consistiu de uma análise dos valores das medidas de interesse e do conteúdo das regras mais bem ranqueadas resultantes deste processo para três bases de dados – *Iris* e *Flags*, obtidas do Repositório de Aprendizado

de Máquina da UCI (Universidade da Califórnia, Irvine – archive.ics.uci.edu/ml (Asuncion e Newman, 2007)), e *Café*, obtida da Embrapa (Empresa Brasileira de Pesquisa Agropecuária), descritas a seguir.

A base *Iris* contém 150 amostras de três espécies de flores do gênero *Iris* (*Iris setosa*, *Iris versicolor* e *Iris virginica*), com cinco atributos no total, descritos na **Tabela 4.1**.

Tabela 4.1: Atributos da base de dados *Iris*.

Atributo	Domínio	Descrição	Valores
petalLength	Decimal	Comprimento da pétala, em cm	[1,0, 6,9]
petalWidth	Decimal	Largura da pétala, em cm	[0,1, 2,5]
sepalLength	Decimal	Comprimento da sépala, em cm	[4,3, 7,9]
sepalWidth	Decimal	Largura da sépala, em cm	[2,0, 4,4]
class	String	Espécie de <i>Iris</i>	{ <i>Iris setosa</i> , <i>Iris versicolor</i> , <i>Iris virginica</i> }

Para o estudo, foram utilizados todos os atributos, que foram pré-processados da seguinte maneira: para os quatro primeiros atributos, o intervalo de valores indicado na tabela anterior foi dividido em três partes de igual tamanho (*small*, *medium* e *large*). Para cada ocorrência dentro dessas faixas de valores, foi atribuído um par atributo-valor, abreviado da seguinte maneira: P foi usado para “*petal*”, S para “*sepal*”, L para “*length*”, W para “*width*”, L para “*large*”, M para “*medium*”, S para “*small*”. Assim, PLM significa “*petal length medium*” e assim por diante. Ao final do pré-processamento, foram obtidos 15 pares atributo-valor.

A base *Flags* contém 30 atributos descrevendo bandeiras nacionais de 194 países e de organizações internacionais. Os atributos são descritos na **Tabela 4.2**.

Para o estudo, foi feita uma seleção de atributos, de modo a diminuir a dimensionalidade da base de dados. Foram selecionados os atributos: *religion*, *stripes*, *red*, *green*, *blue*, *white*, *crosses*, *sunstars* e *icon*. O pré-processamento foi feito da seguinte maneira: para cada valor do atributo *religion*, foi criado um item contendo o par *religion*-<valor>; os atributos *red*, *green*, *blue* e *white*, do tipo lógico, foram transformados em itens, quando presentes na bandeira (correspondente à transação); os demais, *stripes*, *crosses*, *sunstars* e *icon*, embora numéricos, foram transformados em itens quando presentes na bandeira, independentemente da quantidade. Ao final, foram obtidos 16 pares atributo-valor.

A base de dados de *Café* contém atributos e avaliação por especialistas de 11.526 amostras de café, tendo sido utilizada para avaliar uma língua eletrônica que mede a qualidade de bebidas (café, vinho etc.) de acordo com características extraídas de amostras das mesmas (Riul Jr. et al., 2002). Cada entrada da base de dados contém 12 atributos, descritos na **Tabela 4.3**. Foram selecionados 9 dos 12 atributos: *agtron*, *moagem*, *qg*, *classe*, *consistência*, *aroma*, *sabor*, *corpo* e *tipo*. Estes atributos, pré-processados, formam 33 pares atributo-valor.

Tabela 4.2: Atributos da base de dados *Flags*.

Atributo	Domínio	Valores
name	String	Nome do país ou organização
landmass	Numérico	Continente ao qual pertence (1 = North America, 2 = South America, 3 = Europe, 4 = Africa, 5 = Asia, 6 = Oceania)
zone	Numérico	Quadrante geográfico ao qual pertence (1 = Northeast, 2 = Southeast, 3 = Southwest, 4 = Northwest)
area	Numérico	$[0, \infty]$
population	Numérico	$[0, \infty]$
language	Numérico	Idioma oficial (1 = English, 2 = Spanish, 3 = French, 4 = German, 5 = Slavic, 6 = Other Indo-European, 7 = Chinese, 8 = Arabic, 9 = Japanese/Turkish/Finnish/Magyar, 10 = Others)
religion	Numérico	Religião (0 = Catholic, 1 = Other Christian, 2 = Muslim, 3 = Buddhist, 4 = Hindu, 5 = Ethnic, 6 = Marxist, 7 = Others)
bars	Numérico	Número de barras verticais na bandeira
stripes	Numérico	Número de listras horizontais na bandeira
colours	Numérico	Número de cores diferentes na bandeira
red	Lógico	Presença ou não da cor vermelha na bandeira
green	Lógico	Presença ou não da cor verde na bandeira
blue	Lógico	Presença ou não da cor azul na bandeira
gold	Lógico	Presença ou não da cor dourada ou amarela na bandeira
white	Lógico	Presença ou não da cor branca na bandeira
black	Lógico	Presença ou não da cor preta na bandeira
orange	Lógico	Presença ou não da cor laranja ou marrom na bandeira
mainhue	String	Cor predominante na bandeira (empates são decididos pela cor mais acima, depois pela cor ao centro e depois pela cor mais à esquerda)
circles	Numérico	Número de círculos na bandeira
crosses	Numérico	Número de cruzes na bandeira
saltires	Numérico	Número de cruzes na diagonal na bandeira
quarters	Numérico	Número de seções em quadrante na bandeira
sunstars	Numérico	Número de símbolos de sol ou de estrela na bandeira
crescent	Lógico	Presença ou não de um símbolo de lua na bandeira
triangle	Lógico	Presença ou não de triângulos na bandeira
icon	Lógico	Presença ou não de uma imagem não-animada na bandeira
animate	Lógico	Presença ou não de uma imagem animada na bandeira
text	Lógico	Presença ou não de letras ou de escrita (por exemplo, um lema ou slogan) na bandeira
opleft	String	Cor no canto superior à esquerda (indo para a direita para desempates)
botright	String	Cor no canto inferior à direita (indo para a esquerda para desempates)

Tabela 4.3: Atributos da base de dados *Café*.

Atributo	Domínio	Valores
codigo	String	Código identificador da amostra {0.01 a 23.01}
entrada	Data	Data de entrada (dia/mês/ano)
agtron	Numérico	Grau de torra do café $[0, 100]$
moagem	String	Grau de moagem, definida pela forma de preparo {fina, media, desc, ***}
nqg	Decimal	Nota de Qualidade Global $[0.0, 10.0]$
qg	String	Qualidade Global {ruim, razoável, bom, muito bom}
classe	String	Espécie do café {conilon, arábica}
consistencia	String	Consistência do café {rio zona, rio, riada, dura, mole}
aroma	String	Aroma do café {suave, suave/intenso, intenso}
sabor	String	Sabor do café {suave, suave/intenso, intenso}
corpo	String	Corpo da bebida {leve, encorpado ou leve/encorpado}
tipo	String	Tipo do grão {Grão, TM (torrado e moído)}

Após uma etapa de pré-processamento, as bases *Iris*, *Flags* e *Café* resultantes deste processo foram submetidas a uma implementação do algoritmo *Apriori*. Foram utilizados valores arbitrários para o suporte mínimo para cada base de dados: 30 (20%), 20 ($\sim 10\%$)

e 3.467 ($\sim 30\%$) para as bases *Iris*, *Flags* e *Café*, respectivamente. Para a primeira base (*Iris*), foram obtidos 25 *itemsets* de tamanho 2, 15 *itemsets* de tamanho 3 e 6 *itemsets* de tamanho 4. Para a base *Flags*, foram obtidos 37 *itemsets* de tamanho 2, 37 *itemsets* de tamanho 3 e 14 *itemsets* de tamanho 4. Por fim, para a base *Café* foram obtidos 35 *itemsets* de tamanho 2, 58 *itemsets* de tamanho 3, 53 *itemsets* de tamanho 4, 25 *itemsets* de tamanho 5 e 5 *itemsets* de tamanho 6.

Para cada base de dados, regras foram extraídas a partir dos *itemsets* obtidos, e a seguir foram ranqueadas de acordo com uma combinação de medidas de interesse (Suporte, Confiança, *Lift*, *Leverage*, *Certainty Factor*, *Jaccard*, *Gini Index* e *Yule's Q*), sendo selecionadas as r regras mais bem ranqueadas, onde r é o número total de *itemsets* obtidos no passo anterior. Foi feito um *ranking* parcial das regras para cada medida de interesse, sendo que para resolver empates entre duas ou mais regras, foi obtida a média aritmética das posições em cada *ranking* parcial e designado o valor obtido para todas elas, de maneira que o valor do *ranking* parcial não influenciasse no peso da medida no *ranking* final. Daí, calculou-se a média aritmética dos *rankings* para cada regra, de modo a obter uma pontuação global para cada uma delas, que determina o *ranking* final. Este mecanismo é ilustrado na **Tabela 4.4**.

Tabela 4.4: Mecanismo de ranqueamento.

Regra	<i>Ranking</i> Medida 1	<i>Ranking</i> Medida 2	...	<i>Ranking</i> Medida m	<i>Ranking</i> Geral
Regra 1	$Valor_{11}$	$Valor_{12}$	...	$Valor_{1m}$	$\frac{\sum_{i=0}^m Valor_{1i}}{m}$
Regra 2	$Valor_{21}$	$Valor_{22}$	...	$Valor_{2m}$	$\frac{\sum_{i=0}^m Valor_{2i}}{m}$
...	...	...	...	...	...
Regra n	$Valor_{n1}$	$Valor_{n2}$	...	$Valor_{nm}$	$\frac{\sum_{i=0}^m Valor_{ni}}{m}$

Ao observar as t melhores regras, para vários valores de t , tal que t é o número de *itemsets*, e ao observar os *itemsets* responsáveis pela geração destas regras, pode-se notar que muitas regras são geradas por poucos *itemsets*, especialmente as mais bem ranqueadas. Por exemplo, cerca de metade das regras obtidas (47,8%, 44,3%, 49,4% para as bases *Iris*, *Flags* e *Café*, respectivamente) são geradas por apenas 5% dos *itemsets*. Além disso, cerca de 25% dos *itemsets* geram quase todas as regras: 93,5% para a base *Iris*, 94,3% para a base *Flags* e 99,4% para a base *Café*. Por outro lado, a maioria dos *itemsets* não é sequer representada no conjunto de regras obtido, embora possa gerar regras interessantes. Na

Figura 4.1, são mostradas as faixas de *itemsets*, de acordo com a sua posição relativa à quantidade de regras que eles geram.

Figura 4.1: Porcentagem de regras geradas a partir dos melhores *itemsets*.

Tomando como ponto de partida a observação anterior, foi feita uma análise da concentração de regras. Estudando a estrutura das regras, nota-se que alguns *itemsets* têm todas as regras geradas a partir deles incluídas no conjunto de regras após o ranqueamento e a filtragem. Isto se deve ao fato de a maioria das regras provenientes do mesmo *itemset* terem valores parecidos ou o mesmo valor para as medidas de interesse. Do ponto de vista semântico, a presença de muitas regras provenientes de um mesmo *itemset* não acrescenta conhecimento significativo para o minerador em processos exploratórios, já que a existência de uma regra de associação sinaliza que possivelmente outras regras de associação envolvendo o mesmo conjunto de itens tenham também sido geradas. Assim, “esconder” estas regras semanticamente redundantes pode representar uma redução na carga cognitiva do minerador durante a exploração do conjunto de regras, já que neste caso haverá uma sensível diminuição do seu tamanho.

4.2.1 Definições

Para mensurar a preservação da diversidade em conjuntos de regras, foi definida uma medida chamada Cobertura de *Itemsets*, que utiliza a definição de Diversidade de Regras. Ambas são apresentadas a seguir.

Definição 4.8 *Seja um conjunto de regras de associação R , a **Diversidade de Regras** (RD – **Rule Diversity**) é dada pelo tamanho do conjunto S' de *itemsets* frequentes necessários para obter todas as regras deste conjunto, conforme a **Equação 4.1**.*

$$RD(R) = || S' || \quad (4.1)$$

Por exemplo, seja o conjunto de transações mostrado na **Tabela 4.5**. Considerando um valor de suporte mínimo 4, obtém-se o conjunto de *itemsets* freqüentes, dado por $S = \{ \{A, B\}, \{C, D\}, \{C, E\}, \{C, F\}, \{D, E\}, \{D, F\}, \{E, F\}, \{C, D, E\}, \{C, D, F\}, \{C, E, F\}, \{D, E, F\}, \{C, D, E, F\} \}$.

Tabela 4.5: Conjunto de transações.

TID	Itens
01	A B C
02	A B
03	C D E
04	A B E F
05	C D E F
06	A B C D
07	C D E F
08	D E F
09	A B C D E F
10	C D E F

Se forem geradas regras com confiança mínima 80%, obtém-se o conjunto R com 24 regras, mostrado na **Tabela 4.6**. Para obter este conjunto de regras, são necessários 10 *itemsets* freqüentes, contidos no conjunto $S' = \{ \{A, B\}, \{C, D\}, \{D, E\}, \{D, F\}, \{E, F\}, \{C, D, E\}, \{C, D, F\}, \{C, E, F\}, \{D, E, F\}, \{C, D, E, F\} \}$. Assim, a diversidade de regras é $RD(R) = 10$.

Tabela 4.6: Conjunto de regras de associação para confiança mínima 80%.

Antecedente	Consequente	Suporte	Confiança
F	E	6	100,0%
A	B	5	100,0%
B	A	5	100,0%
C,E	D	5	100,0%
C,F	D	4	100,0%
C,F	E	4	100,0%
C,D,F	E	4	100,0%
C,E,F	D	4	100,0%
D,E,F	C	4	100,0%
C,F	D,E	4	100,0%
C	D	6	85,7%
D	C	6	85,7%
D	E	6	85,7%
E	D	6	85,7%
E	F	6	85,7%
F	D	5	83,3%
C,D	E	5	83,3%
D,E	C	5	83,3%
D,F	C	4	80,0%
C,E	F	4	80,0%
D,F	E	4	80,0%
C,D,E	F	4	80,0%
C,E	D,F	4	80,0%
D,F	C,E	4	80,0%

Definição 4.9 Seja um conjunto S de *itemsets* freqüentes obtidos e um conjunto R de regras de associação extraídas, a **Cobertura de Itemsets (IC – Itemset Coverage)**

é uma medida que quantifica a relação entre a diversidade das regras de R e o número de *itemsets* freqüentes contidos em S , dada pela **Equação 4.2**.

$$IC(R, S) = \frac{RD(R)}{\|S\|} \quad (4.2)$$

Para o mesmo conjunto R , e o conjunto de *itemsets* freqüentes S , a cobertura de *itemsets* de R em relação a S é dada por $\frac{10}{12} = 0,833$, ou 83,3%.

O valor da cobertura de *itemsets* fica no intervalo $[0,0, 1,0]$ (0% a 100%), e indica o percentual de representatividade de um conjunto de regras, em relação ao total de *itemsets* freqüentes obtidos após a execução do *Apriori*. Quanto mais próximo de 1,0, melhor é a preservação dos *itemsets* pelo conjunto de regras.

Por outro lado, quanto maior é a quantidade de regras de um conjunto, maior é a probabilidade de seu IC ser alto. Abordagens convencionais de obtenção de regras de associação com filtragem baseada em medidas como suporte e confiança resultam em valores muito baixos de IC quando a quantidade de regras obtidas é pequena, principalmente no caso de grandes bases de dados. Estas abordagens também conseguem, eventualmente, obter percentuais altos de IC , mas para grandes conjuntos de regras, o que dificulta ou inviabiliza a exploração pelo minerador. Assim, uma abordagem eficiente para a obtenção de regras de associação é gerar conjuntos reduzidos de regras que têm, ao mesmo tempo, IC alto, que é uma das contribuições da abordagem I_2E . Seja o mesmo conjunto de transações apresentado anteriormente, um exemplo de conjunto reduzido de regras com alto valor de IC é apresentado na **Tabela 4.7**.

Tabela 4.7: Conjunto reduzido de regras de associação.

Antecedente	Consequente	Suporte	Confiança
A	B	5	100,0%
F	E	6	100,0%
C,E	D	5	100,0%
C,F	D	4	100,0%
C,F	E	4	100,0%
C,D,F	E	4	100,0%
C	D	6	85,7%
D	E	6	85,7%
F	D	5	83,3%
D,F	E	4	80,0%
C	E	5	71,4%
C	F	4	57,1%

Considerando que a última regra foi descartada por ter confiança destacadamente mais baixa que a dos demais, a RD deste conjunto é 11. Desta maneira, o valor da IC para este conjunto com 11 regras é $\frac{11}{12} = 91,7\%$. Este conjunto de regras é menor (11 regras)

do que o anterior (24), e tem uma *IC* (91,7%) maior (83,3%). Portanto, ele tem um poder de representatividade do conhecimento maior, além de ser mais confortável de explorar, comparativamente ao anterior.

Neste trabalho, buscamos uma abordagem interativa que favoreça a geração de conjuntos de regras compactos, mas que consigam preservar a diversidade das regras e a cobertura de *itemsets*. Com isso, pretende-se gerar um conjunto de regras representativas da base de transações, que sirva como ponto de partida para uma exploração pelo usuário do espaço de regras que podem ser geradas por este conjunto de transações.

4.3 Metodologia Interativa para Extração de Regras Representativas

A metodologia interativa para extração de regras engloba as seguintes etapas: (1) execução interativa do Apriori; (2) seleção visual de *itemsets* frequentes pelo minerador; e (3) extração de regras baseada em *itemsets* e orientada por agrupamentos de *itemsets* similares. Estas etapas, e como elas se integram para compor a metodologia interativa, são discutidas nas próximas seções. Elas são apoiadas pelo módulo I₂E, cuja interface de usuário é composta por cinco componentes principais: *ControlPanel*, *ItemsFrame*, *ItemsetsFrame*, *RulesFrame* e *LogFrame*, numeradas de (1) a (5) na **Figura 4.2**. Cada um dos componentes é descrito a seguir.

- (1) *ControlPanel* – Engloba o menu, a barra de ferramentas e o controle de suporte mínimo. Os dois primeiros contêm as funções relacionadas à entrada e à saída (leitura e gravação de arquivos), à navegação pelos resultados do *Apriori* interativo, à configuração da visualização de *itemsets*, geração de regras, configuração e informações sobre o sistema.
- (2) *ItemsFrame* – Apresenta uma lista dos itens contidos nos *itemsets* apresentados na visualização, coordenada com a visualização de *itemsets*, de modo que a seleção de um conjunto de *itemsets* reflete na seleção de itens da lista e vice-versa. Este recurso é uma poderosa ferramenta de interação na exploração dos *itemsets*.
- (3) *ItemsetsFrame* – Contém a visualização de *itemsets*, além de controles para as ferramentas de interação e exploração disponíveis. Entre as ferramentas de interação, podem ser destacadas: o sistema de busca, os controles para ocultação e eliminação de *itemsets*, controles para seleção de escala de cor e do atributo a ser mapeado utilizando cores (se o suporte ou o identificador do agrupamento ao qual pertence

Figura 4.2: Interface do módulo I₂E.

o *itemset*), o filtro que controla a exibição de relacionamentos entre os *itemsets*, e ferramentas para gerenciamento dos agrupamentos de *itemsets* (que pode ser tanto automático quanto manual).

- (4) *RulesFrame* – Apresenta uma lista das regras geradas até o momento e controles de busca e de seleção de regras. A lista pode ser salva em arquivo, para exploração no módulo RuleX.
- (5) *LogFrame* – Mostra as operações realizadas no sistema. O controle de *log* permite registrar todas as operações realizadas. Estas informações são gravadas em um arquivo para estudo posterior à utilização do sistema.

4.3.1 Execução Interativa do Apriori

Em uma execução convencional do *Apriori*, inicialmente são obtidos os *itemsets* frequentes de tamanho 1. A partir daí, são obtidos os *itemsets* frequentes de tamanho 2, 3 e assim sucessivamente até que nenhum *itemset* seja gerado. A metodologia proposta é baseada em uma execução interativa, na qual os passos para obtenção dos *itemsets* frequentes são executados gradualmente, à medida que o usuário for requisitando. O objetivo da execução interativa do algoritmo *Apriori* (Figura 4.3) é permitir que os *itemsets* frequentes gerados

a cada etapa possam ser analisados. O modelo da execução adotado é ilustrado na **Figura 4.3**, e as etapas de extração de *itemsets* são representadas pelo *loop*. Em uma execução automática, os resultados intermediários não são monitorados pelo usuário.

Figura 4.3: Execução interativa do *Apriori*.

A metodologia prevê ainda que o usuário possa navegar pelos resultados de cada passo, avançando ou retrocedendo no processo, bem como alterando parâmetros. A navegação amplia o poder de interferência do minerador no processo, pois ele pode alterar o direcionamento para a realização dos seus objetivos. Na interface do módulo I₂E, a navegação pelos passos pode ser feita por meio do *ControlPanel* – **área (1)** da interface (**Figura 4.2**). Na **Figura 4.4**, são indicadas as principais funções do painel de controle.

Figura 4.4: Painel de controle do Sistema I₂E.

O processo pode ser direcionado de duas maneiras. Uma é alterando o suporte mínimo, aumentando ou diminuindo seu poder de filtragem para gerar mais ou menos *itemsets*.

Isso auxilia o usuário no problema de determinação do suporte, reduzindo o custo de processamento e o tempo de espera. Na interface do módulo I₂E, a alteração do suporte mínimo pode ser feita no *ControlPanel* – área (1) da **Figura 4.2**.

A outra é selecionando *itemsets* de interesse a cada passo. Se *itemsets* freqüentes em uma etapa não são de interesse, a abordagem não gera os *itemsets* freqüentes originados a partir deles em etapas posteriores do processo, diminuindo o número de *itemsets* obtidos.

Para apoiar a execução interativa do *Apriori*, foi implementada uma estrutura *trie*, baseada em uma implementação de Michael Holler¹. A *trie* é uma árvore prefixada (*prefix-tree*) que guarda as informações sobre cada um dos *itemsets* freqüentes, tais como itens que o compõem e freqüência (suporte).

No primeiro nível, a estrutura guarda as informações sobre os *itemsets* de tamanho 1, que são obtidos inicialmente. Em cada nível k da estrutura, são guardadas as informações sobre os *itemsets* freqüentes de tamanho k obtidos no passo k correspondente. A travessia da árvore partindo da raiz até este nível k representa um determinado *itemset* de tamanho k , cujos itens são dados pelo caminho percorrido. É apresentada na **Figura 4.5** uma estrutura *trie* para o *lattice* apresentado no **Capítulo 1**, em que os valores de suporte mínimo e confiança mínima utilizados foram 40,0% e 80,0%, respectivamente.

Figura 4.5: *Lattice* e estrutura *trie* correspondente.

Esta estrutura de dados permite aumentar dinamicamente o limiar de suporte mínimo. Dada uma estrutura *trie* construída para um nível de suporte mínimo s , se um novo nível mínimo s' é dado, tal que $s' \geq s, s > 0$, o algoritmo percorre a árvore procurando pelos *itemsets* freqüentes com suporte s_i que satisfazem a condição $s_i \geq s'$. Quando um *itemset*

¹Obtida de <http://www.helsinki.fi/~holler/datamining/src/>.

freqüente não atende à condição, o algoritmo poda a sub-árvore do *itemset*. Assim, o nível de suporte mínimo pode ser dinamicamente aumentado nas etapas seguintes do *Apriori*.

4.3.2 Exploração Visual de Itemsets Freqüentes

A segunda etapa do processo interativo é a exploração de *itemsets* freqüentes por meio de uma representação visual, ilustrada na **área (3)** da **Figura 4.2**. Em uma execução convencional, não há exploração de *itemsets*. Em abordagens mais recentes, a exploração é feita utilizando visualizações pouco representativas. Nesta metodologia, é apresentada uma visualização que favorece a percepção de relações de (dis-)similaridade entre *itemsets* freqüentes. Este tipo de visualização agrupa visualmente *itemsets* freqüentes que possuem itens em comum, favorecendo a sub-divisão do espaço de busca para a geração e exploração de regras obtidas a partir destes grupos.

A representação utilizada é uma visualização baseada em grafos de *itemsets* freqüentes obtidos em uma etapa específica do *Apriori* (a cada passo, a visualização representa uma camada do *lattice*). No grafo, cada *itemset* corresponde a um vértice de um grafo e é representado por um círculo, cujo tamanho (*CS* – *Circle Size*) mapeia o número de itens do *itemset* freqüente correspondente, segundo a **Equação 4.3**.

$$CS(S_j) = base + size(S_j), \quad (4.3)$$

onde *base* é um valor definido previamente e *size*(S_j) é o tamanho do *itemset*.

O fato de dois *itemsets* freqüentes diferentes estarem conectados por uma aresta indica que eles têm itens em comum. A espessura da linha (*LT* – *Line Thickness*) mapeia o número de itens em comum, e é dada pela **Equação 4.4**.

$$LT(S_j, S_k) = |S_j \cap S_k|, \quad (4.4)$$

onde $|S_j \cap S_k|$ é o número de itens em comum entre S_j e S_k .

O posicionamento dos *itemsets* é determinado por uma técnica de projeção multidimensional, de modo que *itemsets* freqüentes similares, segundo alguma métrica definida pelo usuário, são posicionados mais próximos entre si do que *itemsets* dissimilares. Inicialmente, foram consideradas as técnicas de projeção *IDMAP*, *LSP* e *ProjClus*. A primeira é mais precisa que as demais, porém mais custosa computacionalmente, sendo adequada quando o volume de elementos a serem projetados não é excessivamente grande. A segunda e a terceira são técnicas com menor complexidade computacional e que preservam, na medida do possível, as relações de (dis-)similaridade entre elementos. Por estas

características, estas duas últimas são mais adequadas quando necessita-se de respostas rápidas, e a exigência por precisão é um pouco menor, o que não é o caso da abordagem interativa deste trabalho. Desta maneira, estudos e testes iniciais indicaram que a *IDMAP* é mais adequada para projetar dados desta natureza por preservar melhor as relações de (dis-)similaridade entre os *itemsets*.

Para gerar a projeção dos *itemsets* freqüentes, foram extraídos vetores de características dos *itemsets*. Cada vetor mapeia a ocorrência ou não de um item em um *itemset*, conforme descrito a seguir. Dados um conjunto de *itemsets* $S = \{S_1, S_2, \dots, S_n\}$ e um conjunto de itens $I = \{I_1, I_2, \dots, I_m\}$, tal que $\forall I_i \in I, I_i \subset S_j, S_j \in S$, extrai-se um vetor de características $S_j[]$ de tamanho $|I|$ para cada *itemset* S_j . Para cada item I_i , se $I_i \subset S_j$, $S_j[I_i] = 1$; 0 caso contrário. Então, uma métrica de dissimilaridade adequada, por exemplo, *CityBlock*, é aplicada para obter uma matriz de distâncias (dissimilaridades) entre cada par de *itemsets*. Neste trabalho, foram utilizadas duas métricas de distância: *CityBlock* e Euclidiana (no **Capítulo 5**, é apresentado um estudo da utilização destas duas métricas). A partir da matriz de distâncias, é então gerada uma projeção multidimensional usando a técnica *IDMAP*.

Na projeção obtida, os círculos podem se sobrepor na visualização de *itemsets*. Para diminuir a sobreposição, é empregado um algoritmo de *jitter* para perturbar a posição dos círculos sobrepostos. A visualização é dinâmica, de modo que o posicionamento dos círculos também pode ser mudado manualmente.

A cor de cada círculo pode mapear o suporte do *itemset* correspondente ou o identificador (ID) do agrupamento ao qual ele pertence, de acordo com uma escala de cores selecionada. O suporte é obtido a partir da busca pelo próprio *itemset* freqüente, na estrutura *trie*. Quanto ao agrupamento, o sistema calcula um agrupamento dos *itemsets*, utilizando o algoritmo *k*-médias, no espaço multidimensional ou no espaço bidimensional (como é visto na visualização). O objetivo é auxiliar o minerador na identificação dos agrupamentos visuais já determinados pelo posicionamento. É mostrado na **Figura 4.6** um exemplo de visualização de *itemsets*. São exibidos também os principais elementos presentes no painel de visualização do sistema, como a escala de cores, o histograma e os controles para interação.

Diversas escalas de cores são fornecidas, sendo que a escala em uso é exibida na parte superior esquerda da janela de visualização (ver **Figura 4.6**), e os valores são divididos em 256 partes (de 0 a 255). Todas as escalas de cores implementadas no sistema foram obtidas do *Color Center* (Levkowitz, 1997) do (*IVPR*) *Institute for Visualization*

Figura 4.6: Painel de visualização do Sistema I₂E.

and Perception Research, com exceção da última, que foi obtida da ferramenta *PEx*². São elas: *BTC* – Blue To Cyan (Thrasos Papas, AT&T Bell Labs, Murray Hill, NJ); *BTY* – Blue To Yellow (Canon Research, Sydney, Australia); *Gray* – Non-linearized Grayscale, *LOCS* – Linearized Optimal Color Scale (linearizado perceptualmente) e *OCS* – Optimal Color Scale (Haim Levkowitz); *HeatedObject*, *LinGray* – Linearized Grayscale, *Magenta* e *Rainbow* (todos linearizados perceptualmente) (Steve Pizer, UNC Chapel Hill); e *NewRainbow* (*PEX*). Elas são exibidas na **Figura 4.7**.

A escala de cinzas (*Gray*) é bastante eficiente para preservar a ordem dos valores. *LinGray* é uma versão linearizada da escala de cinzas, que uniformiza a distância (dissimilaridade) percebida entre intensidades adjacentes na escala, uma vez que a percepção da variação de intensidade não é linear (Levkowitz, 1997).

Observações iniciais indicam que escalas de cor que envolvem poucos matizes (um ou dois), por exemplo, *BTC*, *BTY*, *HeatedObject*, *Magenta*, favorecem a visualização do nível de suporte, pois conseguem preservar a ordem dos valores representados para os *itemsets*. O *HeatedObject* é baseado no fato de que o sistema visual humano tem máxima sensibilidade a mudanças de luminância para laranja-amarelo. A última (*Magenta*), por sua vez, é considerada, segundo experimentos relatados na literatura, como a que melhor consegue distinguir mudanças de matiz (Levkowitz, 1997).

²Projection Explorer, desenvolvida no Laboratório de Computação de Alto Desempenho (LCAD) do ICMC/USP.

Figura 4.7: Escalas de cores implementadas no sistema I₂E.

Por outro lado, escalas que utilizam muitos matizes, por exemplo, *LOCS*, *OCS*, *Rainbow*, *NewRainbow*, favorecem a identificação dos diferentes agrupamentos. Neste caso, a preocupação não é com a ordem dos *itemsets*, e sim a diferenciação dos agrupamentos aos quais pertencem. Desta maneira, até sete agrupamentos são facilmente identificáveis utilizando a escala *Rainbow*, bastante popular, que utiliza os matizes do arco-íris (vermelho, laranja, amarelo, verde, azul, índigo, violeta).

A mesma janela que exibe a visualização de *itemsets* no sistema I₂E apresenta também um histograma da distribuição de valores de suporte (ou, alternativamente, da distribuição dos *itemsets* nos diferentes agrupamentos) para auxiliar o usuário a avaliar a adequação do valor do suporte mínimo. Por outro lado, a visualização da distribuição dos grupos de *itemsets* permite analisar os tamanhos dos grupos e decidir se eles devem ou não ser modificados. O histograma é mostrado na parte inferior esquerda da janela de visualização (ver **Figura 4.6**).

O histograma é uma sequência de k barras ($4 \leq k \leq 32$, tal que $k = 2^n$), sendo que cada uma representa uma faixa de valores. Dados um conjunto de *itemsets* $S = \{S_1, S_2, \dots, S_n\}$ e um conjunto de valores $V = \{V_1, V_2, \dots, V_n\}$, tal que V_i é o valor correspondente a S_i para suporte ou para o ID do agrupamento ao qual pertence.

A mesma escala de cores aplicada na visualização de *itemsets* freqüentes é utilizada para colorir o histograma. Desta maneira, é possível estabelecer uma correspondência direta entre os *itemsets* e as barras correspondentes, como se percebe na **área (3)** da **Figura 4.2**.

O Sistema I₂E provê coordenação entre as três visualizações: a lista de itens, o grafo de *itemsets* e o histograma. A lista de itens, ilustrada na **parte (2)** da **Figura 4.2**, é coordenada de maneira bidirecional com a visualização de *itemsets* – a seleção de um conjunto de itens da lista dispara a seleção de *itemsets* que contenham todos os itens da seleção; por outro lado, a seleção de um conjunto de *itemsets* dispara a seleção dos itens neles contidos. A cada seleção de um conjunto de *itemsets*, são mostradas na janela de *log* as freqüências dos itens na seleção. O objetivo da coordenação é facilitar a identificação de grupos de *itemsets*, segundo os itens que contêm.

A coordenação entre a visualização de *itemsets* e o histograma, entretanto, é unidirecional – a seleção de um conjunto de *itemsets* dispara a seleção das barras correspondentes, mas a operação inversa não é possível. A coordenação ajuda a identificar a distribuição dos *itemsets* selecionados – pelo suporte, indicando se eles têm baixo suporte, ou pelo agrupamento, identificando a quais grupos eles pertencem.

O sistema ainda fornece um mecanismo para que o usuário se localize no processo de obtenção de *itemsets* freqüentes: a visualização do passo corrente no contexto, ilustrada no canto superior direito da janela de visualização (ver **Figura 4.6**). O objetivo desta visualização é indicar ao usuário o passo corrente no processo. Inicialmente, quando é gerada a visualização de *itemsets* freqüentes de tamanho 2 (o primeiro que pode ser visualizado), a representação indica um nó com o rótulo 2, representado por um círculo preenchido como o passo 2. À medida que outros passos vão sendo executados, novos nós vão sendo mostrados, de modo que o nó correspondente ao passo corrente é preenchido, enquanto que os outros são mostrados sem preenchimento.

Com as representações visuais, o minerador pode conduzir três tipos de análise. A primeira análise consiste em definir se o valor de suporte mínimo é adequado ou não aos seus objetivos, verificando se os *itemsets* freqüentes de interesse foram gerados. Em caso negativo, ele pode alterar este parâmetro e observar o impacto desta mudança em uma nova visualização.

A segunda análise consiste em fazer comparações de valores de suporte entre *itemsets* freqüentes distintos, por meio de consultas ao suporte dos *itemsets* em questão. A partir dessa análise, o minerador pode ainda iniciar a exploração visual partindo dos *itemsets* que possuem maiores valores de suporte.

A terceira análise consiste em identificar agrupamentos de *itemsets* frequentes. Este tipo de análise permite que o usuário foque a exploração em grupos de *itemsets*, dividindo a atividade de exploração em atividades menores. Além disso, esta análise permite ao usuário inferir correlações entre os itens contidos nos grupos. Por exemplo, dados dois *itemsets* X, Y e X, Z , o usuário pode se interessar em consultar o *itemset* Y, Z para uma análise comparativa.

4.3.3 Filtragem de Regras Baseada em Itemsets

A terceira etapa da metodologia consiste na extração de conjuntos compactos de regras, baseada na obtenção de regras representativas de cada *itemset* frequente, seguida por uma filtragem das regras obtidas orientada por agrupamentos de *itemsets* frequentes geradores das mesmas.

O objetivo é obter conjuntos reduzidos de regras, afetando de maneira mínima a diversidade de *itemsets* frequentes no conjunto. Como as regras obtidas a partir do mesmo *itemset* têm os mesmos itens, variando o antecedente e o conseqüente, a abordagem assume que, em um processo exploratório, nem todas as regras precisam estar explícitas no conjunto de regras a ser explorado: regras geradas a partir do mesmo *itemset* podem ser ocultadas e depois recuperadas, se conveniente. Este fato facilita a exploração porque reduz sensivelmente a quantidade de regras.

Figura 4.8: Processo de filtragem de regras baseada em itemsets.

O processo, ilustrado na **Figura 4.8**, consiste em extrair todas as regras a partir de cada *itemset* e preservar as melhores como representativas. As demais são guardadas

separadamente para a exploração utilizando relações entre regras. A filtragem pode utilizar o **princípio de Pareto**³, segundo o qual, um pequeno número de causas (cerca de 20%) é responsável pela maioria dos fenômenos (cerca de 80%). Uma outra política de filtragem é a **política mínima**, na qual uma única regra por *itemset* é escolhida como representativa. Esta última é utilizada no Sistema I₂E.

Definida a política de filtragem das regras representativas, é possível calcular o **Percentual de Redução Total Máximo** (*MTRP* – *Maximum Total Reduction Percentage*). Dados o tamanho do *itemset* $k \geq 2$, o número de itens presentes na base de dados $n \geq k$ e o número de regras selecionadas pela política de corte (r), o *MTRP* é dado pela **Equação 4.5**.

$$MTRP(k, n) = \left(1 - \frac{r}{\sum_{i=2}^n (C_{n,i} * (2^i - 2))} \right) * 100\% \quad (4.5)$$

É mostrado na **Tabela 4.8** o percentual de redução total do número de regras para diferentes quantidades de itens, considerando a possibilidade de seleção de uma única regra para representar cada *itemset*.

Tabela 4.8: Percentual de Redução Total Máximo de regras, uma regra por *itemset*.

Número de itens (n)	Número potencial total de regras	Uma regra por <i>itemset</i>	Percentual de Redução Total Máximo (MTRP)
2	2	1	50,00%
3	12	4	66,67%
4	50	11	78,00%
5	180	26	85,55%
6	602	57	90,53%
7	1932	120	93,78%
8	6050	247	95,92%
...	...	...	...

As políticas de seleção adotam um **sistema de ranqueamento** das regras segundo um conjunto de medidas de interesse, o qual é ilustrado na **Tabela 4.4**. Dado um conjunto $R = \{r_1, r_2, \dots, r_n\}$ de regras de associação e um conjunto $M = \{m_1, m_2, \dots, m_k\}$ de medidas de interesse, para cada medida de interesse $m \in M$, as regras são ranqueadas em ordem decrescente da medida.

³The Non-Pareto Principle; Mea Culpa – www.juran.com/pdf/SP7518.doc.

Em caso de empate entre duas ou mais regras $r_i, \dots, r_j$, calcula-se o *rank pos* = $\frac{\sum_{p=i}^j r_p}{(j-i)+1}$ para todas as regras $r_i, \dots, r_j$ envolvidas no empate. Feito isso para todas as medidas de interesse, é calculado o *rank* médio para cada regra, o qual determina o *ranking* final.

O conjunto de regras assim obtidas é compacto, porém pode incluir regras que têm valores baixos para as medidas de interesse. É possível eliminar estas regras com uma filtragem adicional determinada por um limiar que elimine as regras pior ranqueadas. Entretanto, se o mesmo limiar absoluto for aplicado ao conjunto de regras como um todo, a filtragem pode se concentrar em poucos grupos de *itemsets* freqüentes, o que pode afetar a variedade do conjunto de regras. Isso é evitado empregando um algoritmo de agrupamento para segmentar os *itemsets* a partir dos quais os grupos de regras são gerados e aplicando uma política de corte dentro de cada um. O algoritmo de agrupamento empregado foi o *k*-médias, por ser bastante simples e apresentar bons resultados.

Em estudos iniciais, foram utilizados três valores arbitrários para a política de filtragem adicional: 15%, 25% e 50%. No sistema I₂E, é indicado por padrão o valor 85% (filtragem de 15%), apesar de ser possível utilizar qualquer valor dentro do intervalo [0-100%]. Na **Seção 5.5** é apresentado um estudo sobre a política de filtragem.

As regras extraídas nesse processo são exibidas na **área (4)** da **Figura 4.2**, podendo ser salvas em arquivo para posterior exploração no módulo RulEx. Um controle de log, ilustrado na **área (5)** da **Figura 4.2** registra as ações do usuário no módulo. Ao final, é obtido um conjunto bastante reduzido de regras, que é usado como ponto de partida para a exploração direcionada de regras.

4.4 Estratégia para Exploração Direcionada de Regras

A estratégia para a exploração direcionada de regras diz respeito à exploração de regras aos pares, com auxílio de uma visualização de medidas de interesse das regras de associação. A exploração é apoiada pelo módulo RulEx, cuja interface com o usuário é composta por cinco partes: *MainPanel*, *ControlPanel*, *RelatedRulesPanel*, *VisFrame* e *FinalRulesPanel*, numeradas de (I) a (V) na **Figura 4.9**. Cada uma delas é descrita a seguir.

- (I) *MainPanel* – Apresenta uma tabela contendo as regras selecionadas como representativas, as quais são usadas como ponto de partida para a exploração de todo o conjunto de regras. Este painel oferece recursos para o usuário filtrar regras por tamanho (em número de itens) e para buscar regras que contém itens especificados.

Figura 4.9: Interface do módulo RuleX.

- (II) *ControlPanel* – Contém controles para filtragem do conjunto de regras representativas segundo diferentes medidas de interesse: Suporte, Confiança, *Lift*, *Leverage*, *Certainty Factor*, *Jaccard*, *Gini Index*, *Yule's Q* e *All Measures*. O usuário pode selecionar uma ou mais medidas de interesse para definir um filtro com base na combinação das medidas selecionadas. Estes controles podem ser ativados ou desativados⁴. Esta ativação é refletida na visualização de regras: são mostradas apenas regras que satisfazem as medidas de interesse ativadas. Por *default*, as medidas Suporte, Confiança, *Lift* e *All measures* são ativadas.
- (III) *RelatedRulesPanel* – Apresenta um conjunto de tabelas, sendo que cada uma delas exibe o resultado de uma relação entre uma regra representativa selecionada no *MainPanel* com outras regras relacionadas, sendo elas representativas ou não. As relações disponíveis no sistema são: *Same Antecedent* (mesmo antecedente), *Same Consequent* (mesmo conseqüente), *Superset* (gerado por um superconjunto do *itemset* gerador), *Subset* (gerado por um subconjunto do *itemset* gerador), e

⁴Na versão 0.11.8, os controles para as medidas Suporte, Confiança e *All measures* não podem ser desativados.

Same Itemset (gerado pelo mesmo *itemset* gerador). Em cada uma destas tabelas, também há filtros por tamanho de regra e um sistema de busca por itens, assim como no *MainPanel*.

- (IV) *VisFrame* – Contém a visualização de medidas de interesse associadas a regras utilizando Coordenadas Paralelas. Cada cor corresponde a uma tabela: a cor vermelha mapeia as regras da *MainPanel* (*Selected rules*). As cores verde, azul, amarelo, magenta e ciano mapeiam, respectivamente, regras (*Related rules*) das tabelas citadas anteriormente: *Same Antecedent*, *Same Consequent*, *Superset*, *Subset*, e *Same Itemset*. A regra é exibida como uma polilinha, cuja cor corresponde à tabela à qual pertence; quando um conjunto de regras é representado, utiliza-se uma região com a mesma cor, porém com um grau de transparência.
- (V) *FinalRulesPanel* – Apresenta uma tabela contendo as regras selecionadas pelo usuário. Estas regras podem ser gravadas em um arquivo para posterior uso na aquisição de conhecimento.

4.4.1 Exploração de Regras por Pares

O conjunto de regras representativas oferece o ponto de partida para a exploração, direcionada pelo usuário, do espaço de regras extraídas. Uma maneira de analisar um conjunto de objetos é por meio de comparação aos pares (no caso, regra com regra ou regra com conjunto de regras), de maneira que características dos objetos são comparadas utilizando alguma métrica, que pode ser subjetiva. A partir dessas comparações, os objetos podem ser ordenados de maneira transitiva, dos mais interessantes até os menos interessantes. A exploração direcionada de regras adota esta abordagem.

O minerador pode comparar pares de regras (regra-regra ou regra-conjunto), guiado por relações entre elas e auxiliado por uma visualização por Coordenadas Paralelas das medidas de interesse associadas.

Primeiramente, seleciona-se uma regra representativa de interesse na **área (I)** da **Figura 4.9**. O usuário pode aplicar os filtros, localizados na **área (II)** da **Figura 4.9** podem ser aplicados à lista da **área (I)**. Depois, seleciona-se uma relação de comparação (na **área (III)** da **Figura 4.9**). Automaticamente, regras que satisfazem a relação são apresentadas ao minerador, que pode compará-las e escolher as regras mais interessantes do seu ponto de vista, descartando as demais.

No escopo deste trabalho, uma relação é definida por uma propriedade, de maneira que regras podem ser ditas relacionadas quando satisfazem esta propriedade. A metodologia

provê cinco relações de comparação: (1) mesmo antecedente, (2) mesmo conseqüente, (3) regra obtida a partir de um superconjunto do *itemset* gerador, (4) regra obtida a partir de um subconjunto do *itemset* gerador e (5) mesmo *itemset*. Para cada relação, é mostrada uma tabela com a lista das regras relacionadas à regra selecionada. Estas relações de comparação são descritas a seguir.

- **Mesmo antecedente da regra em foco** – Dada uma regra de interesse r , pertencente ao conjunto R de regras representativas pré-selecionadas, esta relação obtém um subconjunto S de regras pertencentes ao conjunto $T = \{t_1, t_2, \dots, t_k\}$ e que satisfaçam a relação $antec(r) = antec(t_i)$, tal que $1 \leq i \leq k$. Esta relação permite ao minerador descobrir qual antecedente leva ao conseqüente com melhores valores para as medidas de interesse. Por exemplo, sejam as regras $C \rightarrow E$ (com suporte 5 e confiança 71,4%) e $C \rightarrow F$ (com suporte 4 e confiança 57,1%) do exemplo do **Capítulo 1**. Comparando-as, pode-se inferir que C leva a E com maior intensidade do que leva a F .
- **Mesmo conseqüente da regra em foco** – Dada uma regra de interesse r , pertencente ao conjunto R , esta relação obtém um subconjunto S de regras pertencentes ao conjunto $T = \{t_1, t_2, \dots, t_k\}$ e que satisfazem a relação $conseq(r) = conseq(t_i)$, tal que $1 \leq i \leq k$. Com esta relação, o minerador pode inferir qual o antecedente que leva ao conseqüente da regra com melhores valores de medidas de interesse. Por exemplo, sejam as regras $C \rightarrow E$ (com suporte 5 e confiança 71,4%) e $D \rightarrow E$ (com suporte 6 e confiança 85,7%). Comparando-as, pode-se inferir que D leva a E com maior intensidade, em comparação a C .
- **Gerada por um superconjunto do *itemset* gerador da regra em foco** – Dada uma regra de interesse r , pertencente ao conjunto R , esta relação obtém um subconjunto S de regras pertencentes ao conjunto $T = \{t_1, t_2, \dots, t_k\}$ e que satisfazem a relação $(antec(r) \cup conseq(r)) \subseteq (antec(t_i) \cup conseq(t_i))$, tal que $1 \leq i \leq k$. Esta relação permite ao minerador avaliar como a adição de um item no antecedente (ou conseqüente) altera os valores de medidas de interesse. Por exemplo, sejam as regras $C, F \rightarrow E$ (com suporte 4 e confiança 100%) e $C, D, F \rightarrow E$ (com suporte 4 e confiança 100%). Comparando-as, pode-se inferir que a presença de D no antecedente não alterou os valores das medidas de interesse.
- **Gerada por um subconjunto do *itemset* gerador da regra escolhida** – Dada uma regra de interesse r , pertencente ao conjunto R , esta relação obtém um subconjunto S de regras pertencentes ao conjunto $T = \{t_1, t_2, \dots, t_k\}$ e que satisfazem

a relação $(antec(r) \cup conseq(r)) \supseteq (antec(t_i) \cup conseq(t_i))$, tal que $1 \leq i \leq k$. De maneira semelhante à relação anterior, o minerador pode avaliar como a subtração de um item no antecedente (ou no conseqüente) altera os valores de medidas de interesse. Por exemplo, sejam as regras $C, F \rightarrow E$ (com suporte 4 e confiança 100%) e $C \rightarrow E$ (com suporte 5 e confiança 71,4%). Comparando-as, pode-se inferir que a ausência de F diminuiu a confiança da regra, apesar de ter aumentado o suporte.

- **Mesmo itemset gerador** – Dada uma regra de interesse r , pertencente ao conjunto R , esta relação obtém um subconjunto S de regras pertencentes ao conjunto $T = \{t_1, t_2, \dots, t_k\}$ e que satisfazem a relação $(antec(r) \cup conseq(r)) = (antec(t_i) \cup conseq(t_i))$, tal que $1 \leq i \leq k$. Esta relação permite ao minerador inferir se uma regra tem valores para as medidas de interesse melhores do que as regras relacionadas a ela. O minerador pode inferir que os itens envolvidos têm uma alta correlação se os valores das medidas de interesse forem altos ou inferir que eles estão similarmente distribuídos na base de dados. Por exemplo, sejam as regras $E \rightarrow F$ (com suporte 6 e confiança 85,7%) e $F \rightarrow E$ (com suporte 6 e confiança 100%). Comparando-as, pode-se inferir que E está mais condicionado à ocorrência de F do que o oposto.

4.4.2 Visualização de Medidas de Regras de Associação

Comparar medidas de interesse de duas regras é uma tarefa trivial, apesar de tornar-se eventualmente trabalhosa conforme a quantidade de medidas envolvidas. Fazer o mesmo para conjuntos de regras pode ser ainda mais trabalhoso e improdutivo sem o apoio de representações visuais.

A técnica Coordenadas Paralelas é utilizada para representar medidas de interesse associadas às regras ou ao conjunto de regras em questão, como ilustrado na **área (IV)** da **Figura 4.9**. Seja $M = \{m_1, m_2, \dots, m_i\}$ um conjunto de medidas, de modo que m_{imin} e m_{imax} sejam os seus valores mínimo e máximo. Cada medida é representada por um eixo (paralelo), delimitados por m_{imin} e m_{imax} . Uma regra $r[i]$ contendo um vetor com os valores das medidas é representada como uma polilinha, determinada por pontos que correspondem às posições normalizadas (em relação a m_{imin} e m_{imax}) dos valores de $r[i]$ para o conjunto de medidas M . Analogamente, um conjunto de regras é representado como um polígono fechado, cujos vértices correspondem às posições normalizadas (em relação a m_{imin} e m_{imax}) dos valores mínimo e máximo do intervalo para cada medida do conjunto M . Uma representação esquemática da visualização de regras por Coordenadas Paralelas é apresentada na **Figura 4.10**.

Figura 4.10: Visualização em coordenadas paralelas.

Ao comparar duas regras, é possível identificar qual delas possui o maior valor para cada uma das medidas de interesse – a regra com maior valor intercepta o eixo de uma medida em um ponto mais acima. De maneira análoga, é possível saber o posicionamento de uma regra, em relação a um conjunto de regras, no que diz respeito ao valor para cada uma das medidas de interesse. Se a polilinha da regra intercepta um eixo em um ponto acima ou próximo do limite superior do polígono para uma medida, então a regra tem um valor alto, em relação ao intervalo de valores do conjunto.

Tomando como ponto de partida a comparação visual, a visualização por Coordenadas Paralelas ajuda o minerador a escolher as regras mais interessantes, salvando-as separadamente em um conjunto final de regras (**área (V)** da **Figura 4.9**). Estas regras podem ser organizadas, por exemplo, com um mecanismo de anotação de regras, a fim de serem usadas posteriormente para tomada de decisões.

4.5 Modelo Interativo para a Tarefa de Regras de Associação

O modelo ilustrado na **Figura 4.11** reflete a abordagem interativa apresentada neste capítulo.

O modelo engloba duas grandes etapas, que correspondem aos dois círculos concêntricos e em cujo centro se encontra o minerador. O círculo interno (primeira etapa) corresponde à obtenção de regras de associação, composta pela extração de regras baseada em *itemsets* frequentes e filtragem orientada por agrupamentos (de *itemsets* geradores), enquanto que o círculo externo (segunda etapa) corresponde à exploração de regras de associação.

O objetivo da primeira etapa, iniciada pelo minerador, é facilitar a obtenção de regras de associação significativas por meio de dicas visuais que permitem incorporar

Figura 4.11: Modelo Interativo para Extração de Regras de Associação.

conhecimento pelo minerador. Desta maneira, é possível direcionar o processo, o que pode ajudá-lo a redefinir ou até mesmo definir novos objetivos. Esta estreita interação influencia enormemente o sucesso do processo de descoberta de conhecimento. Esta etapa é dividida em quatro: (1) definição de tarefas; (2) configuração de parâmetros; (3) extração de *itemsets* frequentes; e (4) obtenção de regras.

Na primeira fase, o minerador define as tarefas a serem executadas, baseado nos objetivos definidos inicialmente. A definição das tarefas corresponde à determinação dos *itemsets* frequentes a serem explorados, e à determinação do formato das regras a serem selecionadas como interessantes para exploração. Na segunda fase, o minerador determina os parâmetros de execução do algoritmo. Na terceira, ele pode focar na exploração e seleção de *itemsets* frequentes de interesse, visando atingir os objetivos inicialmente propostos. Por fim, na quarta fase, regras podem ser geradas a partir dos *itemsets* frequentes de interesse ou a partir de todos os *itemsets* frequentes obtidos até então. Estas regras são adicionadas a um conjunto compacto de regras pré-selecionadas, que podem ser exploradas na segunda etapa (círculo externo).

Antes de iniciar o processo de descoberta de conhecimento, o minerador pode ou não ter objetivos determinados. Em caso positivo, as tarefas envolvendo a exploração de *itemsets* freqüentes podem ser executadas. Enquanto explora os *itemsets* freqüentes, o minerador pode redefinir tarefas, por exemplo, se outros itens se revelarem de interesse. A exploração de *itemsets* freqüentes é também útil quando o minerador não tem objetivos definidos *a priori*, o que caracteriza o cenário exploratório. A visualização de *itemsets* freqüentes apoia a definição dos objetivos. Em ambos cenários, o processo todo pode ser repetido, característica ilustrada no ciclo do círculo interno na **Figura 4.11**.

Na segunda etapa, o objetivo é prover mecanismos para facilitar a exploração do universo de regras, permitindo que o minerador utilize seu conhecimento no processo, filtrando regras desinteressantes (óbvias, redundantes ou irrelevantes aos seus objetivos).

A exploração de regras pode ser aplicada em ambos os cenários possíveis – tarefas definidas previamente ou definidas ao longo do processo. Se o minerador tem os objetivos definidos, eles podem ajudar na escolha da relação de regras a ser utilizada. Daí, a análise baseada em comparações visuais de pares de regras pode confirmar ou descartar as hipóteses levantadas. Por outro lado, se o minerador não definiu os objetivos, então ele pode analisar as regras pré-selecionadas, por exemplo, inspecionando individualmente cada regra, selecionar e comparar as regras relacionadas, segundo suas medidas de interesse. Ao fazer isto, o minerador pode formular hipóteses, que depois podem vir a ser confirmadas ou descartadas.

Durante a extração de regras, o minerador pode redefinir ou formular tarefas novas. Além disso, durante a exploração, as regras obtidas até então podem não ser suficientes. O modelo permite que o minerador lide com ambas situações (setas em tracejado na **Figura 4.11**). Esta característica é fundamental, e diferencia o modelo das abordagens existentes: pelo fato de induzir o minerador a participar ativamente e adquirir informações durante o processo, ele pode focar nos seus objetivos, obter resultados mais ricos (mais regras são compreendidas) e reduzir o tempo de análise (o algoritmo é executado menos vezes, e com melhor aproveitamento).

4.6 Considerações Finais

Neste capítulo, foi apresentada uma abordagem interativa para a realização de tarefas de obtenção de regras de associação, composta por: uma metodologia interativa para extração de regras; e uma estratégia para exploração direcionada de regras. A metodologia é composta por três etapas: (1) execução interativa do *Apriori*, com o intuito de permitir que os resultados intermediários do processo sejam monitorados pelo usuário que pode,

conseqüentemente, direcioná-lo de acordo com seus objetivos; (2) seleção de *itemsets* freqüentes de interesse, apoiada por uma representação visual baseada em grafo cujo posicionamento dos *itemsets* é determinado por uma técnica de projeção multidimensional, com base na similaridade dos *itemsets*, para auxiliar o usuário nesse direcionamento; e (3) extração de regras baseada em *itemsets* e orientada por agrupamentos de *itemsets* similares, com o objetivo de obter conjuntos de regras de associação reduzidos e com bons valores de cobertura de *itemsets*; A estratégia é constituída por uma exploração de regras aos pares, visando a permitir que o usuário avalie o interesse das regras por meio de comparações entre elas, por meio de comparação de medidas de regras de associação, apoiada por uma representação visual baseada em projeção geométrica.

A abordagem pressupõe um modelo para a realização de tarefas de obtenção de regras de associação que considera dois cenários iniciais possíveis: (1) o usuário tem os objetivos bem definidos ao iniciar o processo; e (2) um cenário exploratório em que o usuário não tem os objetivos definidos *a priori*. No primeiro cenário, o usuário utiliza a metodologia para verificar as hipóteses iniciais. Neste caso, durante este processo, o usuário pode redefinir suas hipóteses baseado na exploração do universo de regras, apoiado por representações visuais. No segundo, o usuário utiliza a metodologia para explorar o conjunto de dados e as regras de associação, a fim de levantar hipóteses.

No próximos dois capítulos, são apresentados os resultados de estudos conduzidos para avaliar diversos aspectos da abordagem interativa proposta.

Avaliação da Abordagem: Efeito dos Parâmetros no Processo

5.1 Considerações Iniciais

Para avaliar a abordagem proposta, foram realizadas diversas análises apoiadas pelo Sistema I₂E, com o objetivo de: (1) comparar a abordagem interativa sob diversos aspectos, em relação a uma abordagem convencional de obtenção de regras de associação; e (2) avaliar o efeito de variar alguns parâmetros de execução nos resultados do processo.

Alguns trabalhos relatados na literatura discutem a dificuldade de avaliar técnicas de visualização de informação para análise de dados (Chen e Yu, 2000; Plaisant, 2004). Dificuldades similares se aplicam à mineração visual de dados, que emprega estas técnicas. A avaliação é desafiadora por diversas razões. Primeiro, uma nova abordagem pode ajudar a identificar padrões que nenhuma técnica existente é capaz de identificar, de maneira que uma comparação direta não é possível. Mesmo quando uma comparação é possível, deve-se considerar que o usuário, embutido no processo, pode contribuir com um viés que pode distorcer os resultados, por exemplo, diferentes níveis de proficiência do usuário na tarefa de análise podem invalidar o estudo comparativo. Além disso, a descoberta não é um evento instantâneo, e experimentos controlados devem ter rígidos limites de tempo que limitam o seu escopo. Outras dificuldades incluem a medição de aspectos subjetivos

do usuário no processo, por meio de métricas subjetivas que são susceptíveis a erro, e avaliar o efeito combinado de visualizações. Mesmo que todas estas questões tenham sido resolvidas, os estudos experimentais devem ainda ser estatisticamente válidos para produzir resultados úteis e confiáveis. Por fim, existe ainda a discussão sobre a validade de estudos experimentais controlados para avaliar atividades como a análise de dados e a exploração, que estão longe de serem consideradas controladas em situações reais.

Tomando como referência este contexto, foram conduzidos diversos estudos, quantitativos, para comparar alguns aspectos da abordagem (que podem ser mensurados) com outras. Primeiramente, foi feito um estudo comparativo da cobertura de *itemsets* de conjuntos de regras obtidas com o uso da metodologia, e conjuntos de regras obtidas utilizando uma abordagem convencional, descrito na **Seção 5.2**. Também foi feita uma avaliação da política de filtragem da metodologia, em comparação com uma abordagem convencional e com uma abordagem utilizando *itemsets* freqüentes fechados, apresentada na **Seção 5.3**. Em seguida, avaliou-se a visualização de *itemsets* gerada por projeção multidimensional utilizada, uma vez que diversas visualizações alternativas poderiam ser utilizadas com o mesmo objetivo, apresentada na **Seção 5.4**. Uma avaliação do limiar de filtragem orientada pelos agrupamentos de *itemsets* é apresentada na **Seção 5.5**.

Para as avaliações apresentadas neste capítulo, foram selecionadas diversas bases de dados com características diversas. Três delas (*Iris*, *Flags* e *Café*) foram apresentadas anteriormente. Outras três (PMGRN, Orkut e Corpus Artificial) são apresentadas oportunamente. Um resumo com as características dessas bases é mostrado na **Tabela 5.1**.

Tabela 5.1: Bases de dados utilizadas.

Base de dados	# registros	# atributos	# atributos utilizados	# pares atributo-valor	Valores desconhecidos?
Iris	150	5	5	15	não
Flags	194	30	9	16	não
Café	11.526	12	9	33	sim
PMGRN fêmeas	3.739	23	17	67	sim
PMGRN machos	19.735	20	15	86	não
Corpus Artificial	1.000	128	128	128	não
Orkut	1.606	3.191	3.191	3.191	não

5.2 Avaliação da Cobertura de Itemsets

Este estudo teve como objetivo analisar como se compara a política de extração de conjuntos compactos de regras de associação proposta, com uma abordagem convencional de extração de regras seguida de filtragem com base em medidas de interesse.

Neste estudo, foram utilizadas três bases de dados: *Iris*, *Flags* e *Café*. Para cada uma delas, foram obtidos conjuntos de regras usando uma abordagem convencional com filtragem *a posteriori* e conjuntos de regras usando a metodologia interativa para extração de regras proposta. Inicialmente, foi utilizada a base *Iris*. Foram extraídos conjuntos de regras de associação da base pré-processada com o suporte determinado arbitrariamente em 30 (20%). O primeiro conjunto, usando uma abordagem convencional, será chamado **Conjunto Co**, do inglês *Conventional*, enquanto que o último, usando a metodologia proposta, será chamado **Conjunto I₂E**.

Para obter o **Conjunto I₂E**, foi extraída uma única regra por *itemset*, sendo esta a regra mais bem ranqueada entre as derivadas do *itemset*, de acordo com uma combinação de oito medidas de interesse: Suporte, Confiança, *Lift*, *Leverage*, *Certainty Factor*, *Gini Index*, *Jaccard* e *Yule's Q*. Para obter o **Conjunto Co** foi utilizada uma abordagem convencional sem restrição por medidas de interesse além do suporte (como inicialmente definido), seguida de um ranqueamento das regras, segundo a mesma combinação de medidas de interesse, e seguida por uma filtragem *a posteriori*. Para efeito de comparação, a filtragem foi dirigida de modo que o **Conjunto Co** tivesse o mesmo tamanho do **Conjunto I₂E** com o qual ele deveria ser comparado.

Para medir a diversidade das regras de cada um dos conjuntos (**Conjunto I₂E** e **Conjunto Co**), calculou-se o número de *itemsets* distintos necessários para gerar cada um deles. Desta maneira, comparou-se a diversidade das regras em ambos conjuntos, levando-se em conta o percentual de perda de representatividade dos *itemsets*, isto é, a **cobertura de *itemsets*** (Definição 4.9).

O mesmo procedimento foi adotado para as outras duas bases de dados, *Flags* e *Café*, utilizando, respectivamente, os valores 20 (10,3%) e 3.467 (30,1%) para o suporte, também definidos arbitrariamente.

Os resultados da medição da cobertura de *itemsets* para a base de dados *Iris* são mostrados na **Tabela 5.2**. Como se pode observar, o **Conjunto I₂E** tem uma cobertura de 100% dos *itemsets*, já que uma regra representativa foi selecionada para cada um dos *itemsets* extraídos. Por outro lado, o **Conjunto Co** de mesmo tamanho tem uma cobertura baixa, de apenas 32,6%. Em outras palavras, o conjunto de regras extraído usando uma abordagem convencional seguida por filtragem baseada em um ranqueamento inclui regras obtidas a partir de apenas 15 dos 46 *itemsets*.

Analisando o conjunto de regras, pode-se explicar a baixa cobertura de *itemsets* do **Conjunto Co**. Verificando o *itemset* a partir do qual cada regra foi obtida, nota-se que um conjunto reduzido de *itemsets* é responsável por várias regras geradas. Nesta análise, o *itemset* “*Iris-setosa*, *petalLengthSmall*, *petalWidthSmall*, *sepalLengthSmall*” gerou 12 das

Tabela 5.2: Cobertura de *itemsets*: base *Iris*.

k	Número de <i>itemsets</i> (Número máximo de regras)	Conjunto Co		Conjunto I ₂ E	
		Número de regras (20,53%)	Cobertura de <i>itemsets</i>	Número de regras (20,53%)	Cobertura de <i>itemsets</i>
2	25 (50)	9	6 (24,0%)	25	25 (100,0%)
3	15 (90)	22	6 (40,0%)	15	15 (100,0%)
4	6 (84)	15	3 (50,0%)	6	6 (100,0%)
Total	46 (224)	46	15 (32,6%)	46	46 (100,0%)

46 regras; outro *itemset* gerou seis das 46 regras; três *itemsets* geraram quatro regras cada um; um *itemset* gerou três regras; quatro *itemsets* geraram duas regras cada um; e cinco *itemsets* geraram uma regra cada um. Por outro lado, 22 *itemsets* não são representados no conjunto de regras.

Foram obtidos resultados similares para as bases de dados *Flags* e *Café*, mostrados respectivamente nas **Tabelas 5.3** e **5.4**. Isto confirma a suposição inicial deste trabalho, de que a extração e filtragem de regras pela abordagem convencional falha na preservação da cobertura de *itemsets*.

Tabela 5.3: Cobertura de *itemsets*: base *Flags*.

k	Número de <i>itemsets</i> (Número máximo de regras)	Conjunto Co		Conjunto I ₂ E	
		Número de regras (17,89%)	Cobertura de <i>itemsets</i>	Número de regras (17,89%)	Cobertura de <i>itemsets</i>
2	37 (74)	9	7 (18,9%)	37	37 (100,0%)
3	37 (222)	33	12 (32,4%)	37	37 (100,0%)
4	14 (196)	46	8 (57,2%)	14	14 (100,0%)
Total	88 (492)	88	27 (30,7%)	88	88 (100,0%)

Tabela 5.4: Cobertura de *itemsets*: base *Café*.

k	Número de <i>itemsets</i> (Número máximo de regras)	Conjunto Co		Conjunto I ₂ E	
		Número de regras (7,92%)	Cobertura de <i>itemsets</i>	Número de regras (7,92%)	Cobertura de <i>itemsets</i>
2	35 (70)	6	3 (8,6%)	35	35 (100,0%)
3	58 (348)	30	14 (24,1%)	58	58 (100,0%)
4	53 (742)	62	16 (25,8%)	53	53 (100,0%)
5	25 (750)	62	10 (40,0%)	25	25 (100,0%)
6	5 (310)	16	2 (9,1%)	22	22 (100,0%)
Total	176 (2.220)	176	45 (25,6%)	176	176 (100,0%)

Os **Conjuntos I₂E** têm uma cobertura de *itemsets* maximizada pelo próprio processo de extração, o que ajuda a reduzir a redundância semântica entre as regras. Devido ao fato de estes conjuntos incluírem regras provenientes de diversos *itemsets*, é provável que algumas delas tenham pouco interesse para o minerador, que pode eliminá-las ao seu critério. Para avaliar este aspecto, foram realizadas filtrações adicionais (baseadas no mesmo processo de ranqueamento) para eliminar as regras pior ranqueadas. Inicialmente,

foram utilizados, de maneira arbitrária, como parâmetro de eliminação 15%, 25% e 50% do tamanho do conjunto de regras obtidas. Para as três bases de dados consideradas, analisou-se o impacto dessas filtrações adicionais na cobertura de *itemsets*. Os resultados foram depois comparados com os **Conjuntos Co**, submetidos ao mesmo processo de filtração. Os resultados são mostrados na **Tabela 5.5**.

Tabela 5.5: Cobertura de *itemsets* após a filtração.

Limiar de filtração (%)	<i>Iris</i>		<i>Flags</i>		<i>Coffee</i>	
	Conjunto I ₂ E	Conjunto Co	Conjunto I ₂ E	Conjunto Co	Conjunto I ₂ E	Conjunto Co
0 (sem filtração)	100,0%	32,6%	100,0%	30,7%	100,0%	25,6%
15	84,8%	23,9%	84,1%	28,4%	84,7%	23,3%
25	73,9%	21,7%	73,9%	28,4%	74,4%	21,0%
50	47,8%	17,4%	48,9%	19,2%	49,4%	17,1%

Para a base de dados *Iris*, a filtração de 25% das regras mais mal-ranqueadas eliminou as regras mostradas na **Tabela 5.6**. Todas as regras eliminadas têm valores baixos para as medidas de interesse, principalmente as duas últimas. Sabendo que o interesse de uma regra é subjetivo, a determinação de um limiar para a filtração depende da aplicação e do usuário. Para este conjunto de regras, 25% pareceu um valor razoável para um filtro automático. Na **Seção 5.5** é apresentado um estudo sobre a determinação do limiar de filtração.

Tabela 5.6: Regras eliminadas após a filtração de 25% das regras: base *Iris*.

Antecedente	Consequente	Suporte	Confiança	<i>Lift</i>	<i>Leverage</i>	<i>Certainty Factor</i>	<i>Jaccard</i>	<i>Gini Index</i>
PLS	SWM	36	0,72	1,23	0,04	0,33	0,35	0,02
PWS	SWM	36	0,72	1,23	0,04	0,33	0,35	0,02
setosa	SWM	36	0,72	1,23	0,04	0,33	0,35	0,02
PWL	SLM	30	0,65	1,38	0,06	0,34	0,34	0,03
virginica	SLM	32	0,64	1,35	0,06	0,32	0,36	0,03
SLS	SWM	37	0,63	1,07	0,02	0,10	0,34	0,00
SWM	SLM	37	0,42	0,89	-0,03	-0,10	0,30	0,01

Para a base de dados *Flags*, a filtração de 25% eliminou regras com valores baixos para *Leverage* ($\leq 0,07$), *Jaccard* ($\leq 0,35$) e *Gini Index* ($\leq 0,04$). Por fim, para a base de dados *Café*, a mesma filtração eliminou regras com valores baixos para *Lift* ($\leq 1,17$), *Leverage* ($\leq 0,06$), *Jaccard* ($\leq 0,56$) e *Gini Index* ($\leq 0,03$).

Observando os conjuntos de regras de associação obtidos com as duas abordagens, a convencional com filtração *a posteriori*, e a abordagem proposta, é possível afirmar que a aplicação da política de extração de regras da metodologia é mais adequada quanto à cobertura de *itemsets* do que uma abordagem convencional, considerando conjuntos de mesmo tamanho. Desta maneira, conjuntos de regras obtidos com a abordagem proposta podem representar uma maior quantidade de informação não-redundante no tocante à

diversidade de itens contidos nas regras do que uma abordagem convencional. Além disso, a possibilidade de aplicar uma filtragem adicional orientada por *itemsets*, ou por agrupamentos de *itemsets* semelhantes, evita que sejam obtidas regras com baixos valores para as medidas de interesse.

5.3 Avaliação da Política de Filtragem

Este estudo teve como objetivo comparar os conjuntos de regras de associação obtidos com a metodologia proposta (**Conjuntos I₂E**) com aqueles obtidos por uma abordagem convencional (**Conjuntos Co**) e com o universo de regras possíveis de serem obtidas – *powerset* (**Conjuntos Po**) quanto ao tamanho – em outras palavras, o percentual de redução dos dois primeiros conjuntos em relação ao último, ou o quanto se perde com a filtragem embutida em ambas as abordagens. Também foi feito um estudo do posicionamento relativo das regras dos **Conjuntos I₂E** nos **Conjuntos Co**, buscando analisar a distribuição das regras obtidas com a abordagem proposta, relativamente ao conjunto de regras obtido em uma abordagem convencional.

Neste estudo, foram utilizados os conjuntos de regras de associação selecionados pelo especialista (no segundo estudo com especialista, **Seção 6.2.2**), como sendo as mais interessantes, extraídas da base de dados do PMGRN. A base de dados do Programa de Melhoramento Genético da Raça Nelore (PMGRN) (Lôbo et al., 2008), com 3.739 registros, é resultado de um programa que registra características de bovinos da raça Nelore de diversas partes do país, com o intuito de investigar tecnologias para aumentar a produtividade na criação de bovinos. Desta maneira, são coletadas informações sobre a Diferença Esperada de Progênese (DEP) em diversas características, quantificando a diferença de cada animal em relação a um valor de referência listadas adiante, além de diversas outras informações. São mostrados na tabela **Tabela 5.7** todos os atributos, sendo que os últimos 11 correspondem às DEPs.

A base de dados do PMGRN, como é possível notar, já foi pré-processada por um especialista do domínio. A mudança mais significativa foi nos valores para as DEPs, que no estado bruto são dados em intervalos de valores decimais. No pré-processamento, foram determinados três intervalos de valores para cada uma das DEPs, o que resultou em três valores possíveis para cada atributo: BOTTOM_50, TOP_50 e TOP_25, mostrados na **Tabela 5.7**. Assim, os animais foram separados em faixas correspondendo à sua função no melhoramento: os TOP_25% melhoram a genética, os TOP_50% não ajudam nem atrapalham, enquanto que os BOTTOM_50% pioram a genética.

Tabela 5.7: Atributos da base de dados PMGRN.

Atributo	Domínio	Descrição	Valores
CGA	String	Identificação	{A*}
PALCGA	String	Identificação do pai	{A*}
MAE_CGA	String	Identificação da mãe	{A*}
SX	String	Sexo	{Macho, Femea}
CAT	String	Categoria	{PO (registro genealógico completo), PI (puro de origem, de ascendência importada), LA (registro genealógico incompleto), CL (sem registro genealógico)}
VAR	String	Variedade	{M (mocho – sem chifres), P (padrão – com chifres)}
USA Situacao	String	Última Situação do Animal	{Ativo, Descarte, Morto, Venda}
USA Classe	String	Classe	{Bezerra, Multipara, Nao Identificado, Novilha, Primipara, Touro Erado, Touro Jovem}
USA Ano	Númerico	Ano da Última Situação do Animal	[1971, 2006]
USA Mes	String	Mês da Última Situação do Animal	{Janeiro, Fevereiro, Março, Abril, Maio, Junho, Julho, Agosto, Setembro, Outubro, Novembro, Dezembro}
NFA	String	Identificação do criador	{A, B, C, D, E, F, G, H, I, J}
UF	String	Unidade da Federação	{GO, MG, MS, MT, PA, SP, TO}
REGIAO	String	Região da unidade da federação	{Centro-Oeste, Norte, Sudeste}
F	String	Cosangüineidade	{Menor_10, Zero, Maior_igual_10}
MGT	String	Mérito Genético Total	{MGT_BOTTOM_50, MGT_TOP_50, MGT_TOP_25 Nao_Avaliado}
MP120	String	Habilidade materna	{MP120_BOTTOM_50, MP120_TOP_50, MP120_TOP_25}
DP120	String	Peso aos 120 dias	{DP120_BOTTOM_50, DP120_TOP_50, DP120_TOP_25}
DP365	String	Peso aos 365 dias	{DP365_BOTTOM_50, DP365_TOP_50, DP365_TOP_25}
DPE365	String	Perímetro escrotal aos 365 dias	{DP365_BOTTOM_50, DP365_TOP_50, DP365_TOP_25}
DPAC	String	Produção acumulada	{DPAC_BOTTOM_50, DPAC_TOP_50, DPAC_TOP_25}
DIPP	String	Idade ao primeiro parto	{DIPP_BOTTOM_50, DIPP_TOP_50, DIPP_TOP_25}
DSTAY	String	Probabilidade de permanência no rebanho	{DSTAY_BOTTOM_50, DSTAY_TOP_50, DSTAY_TOP_25 Nao_Avaliado}
D3P	String	Probabilidade de Prenhez Precoce	{D3P_BOTTOM_50, D3P_TOP_50, D3P_TOP_25, Nao_Avaliado}
DAOL	String	Área de Olho de Lombo	{DAOL_BOTTOM_50, DAOL_TOP_50, DAOL_TOP_25, Nao_Avaliado}
DEG	String	Espessura de Gordura na Costela	{DEG_BOTTOM_50, DEG_TOP_50, DEG_TOP_25, Nao_Avaliado}

Para extrair regras de associação dessa base, foi preciso fazer uma codificação específica, criando um item para cada par atributo-valor. Desta maneira, para o valor PO do atributo CAT, é criado um item CAT-PO que combina o par.

Neste estudo, foi utilizado um conjunto de tarefas definido pelo especialista. Para executar cada uma das tarefas, foi extraído um conjunto de regras de associação. A seguir, foi feita uma avaliação quantitativa do número de regras do conjunto final extraído para cada tarefa, em relação ao conjunto equivalente de regras de associação obtida por uma abordagem convencional e em relação ao *powerset* (todas as regras possíveis), dados os mesmos valores para medidas de interesse determinadas. As tarefas são listadas a seguir.

- (i) Identificar perfis de animais para as fazendas (regras envolvendo fazenda e DEP)
- (ii) Identificar perfis de animais para as fazendas (regras envolvendo fazenda e 2 DEPs)

- (iii) Identificar perfis de animais por variedade para as fazendas (fazenda, DEP e variedade)
- (iv) Identificar perfis de animais por categoria para as fazendas (fazenda, DEP e categoria)
- (v) Identificar evolução da qualidade dos animais ao longo do tempo (ano e DEP)
- (vi) Identificar períodos de maior venda de animais por qualidade (mês e DEP)
- (vii) Identificar períodos de maior venda de animais por fazenda (mês e fazenda)

Inicialmente, para as quatro primeiras tarefas – (i), (ii), (iii) e (iv) – foi usado o valor 583 para suporte mínimo. Entretanto, para a tarefa (i), este valor não foi suficiente para extrair regras interessantes, sendo determinado 292 como um valor mais adequado. Da mesma maneira, foi determinado um novo valor para a tarefa (iv), que foi 195. Para as demais – (v), (vi) e (vii) – foram definidos os valores de suporte 308, 199 e 99, respectivamente. Na **Tabela 5.8**, são mostradas as características principais (número de regras, suporte mínimo e confiança mínima do conjunto) dos conjuntos de regras de associação obtidos ao executar estas tarefas. O número de *itemsets* frequentes para cada configuração, detalhado de acordo com o tamanho, é mostrado na **Tabela 5.9**.

Tabela 5.8: Características dos conjuntos de regras selecionadas, por tarefa.

Número da tarefa	Tamanho (Número de regras)	Suporte mínimo do conjunto $[0, \infty]$	Confiança mínima do conjunto $[0,0, 1,0]$
(i)	4	293	0,333
(ii)	13	615	0,315
(iii)	7	586	0,362
(iv)	19	208	0,334
(v)	21	308	0,335
(vi)	67	199	0,177
(vii)	17	99	0,126

Estes conjuntos de regras obtidos com a abordagem interativa foram comparados com conjuntos de regras obtidos com uma abordagem convencional, utilizando os mesmos parâmetros de suporte e de confiança, e com uma abordagem que emprega *itemsets* frequentes fechados, com relação ao tamanho.

5.3.1 Redução do número de regras

A partir dos conjuntos de *itemsets* frequentes obtidos, regras foram geradas com a metodologia interativa para extração de regras apresentada, adotando a política de geração de regras

Tabela 5.9: Conjuntos de *itemsets* freqüentes obtidos, por tarefa.

Tamanho dos <i>itemsets</i>	Tarefa						
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
1	47	40	41	54	45	55	58
2	582	249	273	731	551	748	951
3	1.770	394	450	3.016	1.619	3.169	6.047
4	2.020	215	275	4.666	1.786	5.099	14.859
5	1.223	41	56	3.913	1.037	4.605	18.231
6	353	5	10	2.343	293	3.184	14.041
7	45	–	1	1.183	41	2.075	8.248
8	3	–	–	568	3	1.326	4.346
9	–	–	–	230	–	716	2.228
10	–	–	–	67	–	286	1.024
11	–	–	–	12	–	78	365
12	–	–	–	1	–	13	91
13	–	–	–	–	–	1	14
14	–	–	–	–	–	–	1

uma-regra-por-*itemset*, com uma abordagem de extração que emprega *Frequent Closed Itemsets* e com uma abordagem convencional, utilizando as medidas de interesse definidas para suporte mínimo e confiança mínima. Por fim, os tamanhos dos conjuntos de regras foram comparados com o tamanho do conjunto total de regras possíveis (tamanhos dos conjuntos de regras e percentual de redução em relação ao *Powerset*). Os resultados são mostrados na **Tabela 5.10**. Um asterisco (*) denota as configurações em que a execução foi interrompida antes do término. Dois asteriscos (**) denotam as configurações em que a execução não foi feita. Três asteriscos (***) denotam percentuais não determinados.

Tabela 5.10: Número de regras, em relação ao *powerset*.

Tarefa	Metodologia, com o Sistema I ₂ E		Abordagem com <i>Frequent Closed Itemsets</i>		Abordagem convencional com suporte e confiança		<i>Powerset</i>
	Número de regras	Percentual de redução	Número de regras	Percentual de redução	Número de regras	Percentual de redução	Número de regras
(i)	6.043	98,04%	47.456	54,83%	52.119	50,40%	105.072
(ii)	944	87,26%	4.616	37,72%	5.334	28,04%	7.412
(iii)	1.106	88,38%	5.225	45,13%	7.894	17,10%	9.522
(iv)	16.784	98,04%	*80.063+	***	*132.706+	***	855.288
(v)	5.375	94,09%	39.979	56,08%	45.714	49,78%	91.024
(vi)	21.355	98,88%	**	***	*124.000+	***	1.904.160
(vii)	70.504	99,03%	**	***	**	***	7.240.008

Os resultados mostrados na **Tabela 5.10** comprovam a eficácia na redução do tamanho dos conjuntos de regras obtidos utilizando a metodologia proposta. Para todas as tarefas realizadas, o percentual de redução em relação ao *powerset* ficou entre 87,26% e 99,03%. Usando a metodologia, obtém-se percentuais de redução claramente melhores que as das outras duas abordagens: ao utilizar uma abordagem convencional, é evidenciada uma redução no número de regras, porém o percentual é menor – variou de 17,10% a 50,40%;

empregando o conceito de *Frequent Closed Itemsets*, observou-se uma redução de 37,72% a 56,08%.

Analisando o comportamento da redução do número de regras à medida que o suporte é diminuído (e o número de *itemsets* freqüentes, aumentado) para os conjuntos obtidos com a metodologia proposta, pôde ser observada uma tendência: o percentual de redução aumenta à medida que o suporte mínimo diminui. Este fato pode ser explicado pela maior quantidade de *itemsets* freqüentes de maior tamanho, cujo percentual parcial de redução é maior. Para a abordagem por *Frequent Closed Itemsets*, isto também ocorre, porém a causa é diferente: quanto menor o suporte dos *itemsets*, maior é a sua concentração, sendo maior a probabilidade de haver descarte de *itemsets* freqüentes não-fechados. Para os conjuntos obtidos com a abordagem convencional, observações análogas não se aplicam, pois estes conjuntos foram determinados pela confiança definida arbitrariamente pelo especialista do domínio.

Aplicando-se filtragem adicional e a política regra-por-*itemset* para geração de regras, a redução no número de regras para os conjuntos obtidos com a metodologia é ainda maior, como pode ser visto na **Tabela 5.11**. Utilizando 50% de filtragem, obteve-se uma redução de até aproximadamente 99,51% no número de regras. Isto foi observado para cenários em que o mínimo suporte utilizado era bastante baixo, como para a tarefa (vii). Para qualquer configuração de filtragem adicional, observa-se também o aumento do percentual de redução à medida que o suporte mínimo é diminuído.

Tabela 5.11: Número de regras com a abordagem I_2E , em relação ao *powerset*.

Tarefa	I_2E , 15% filtragem		I_2E , 25% filtragem		I_2E , 50% filtragem		<i>Powerset</i>
	Número de regras	Percentual de redução	Número de regras	Percentual de redução	Número de regras	Percentual de redução	Número de regras
(i)	~5.137	~95,11%	~4.543	~95,68%	~3.022	~97,12%	105.072
(ii)	~802	~89,18%	~708	~90,45%	~472	~93,63%	7.412
(iii)	~940	~90,13%	~830	~91,28%	~553	~94,19%	9.522
(iv)	~14.266	~98,33%	~12.588	~98,53%	~8.392	~99,02%	855.288
(v)	~4.569	~94,98%	~4.031	~95,57%	~2.688	~97,05%	91.024
(vi)	~18.152	~99,05%	~16.016	~99,16%	~10.678	~99,44%	1.904.160
(vii)	~59.928	~99,17%	~52.878	~99,27%	~35.252	~99,51%	7.240.008

5.3.2 Distribuição das regras em relação ao ranking

Sabe-se que as regras de associação obtidas com a abordagem (**Conjuntos I_2E**) são um subconjunto das regras de associação obtidas com uma abordagem convencional (**Conjuntos Co**), que por sua vez são um subconjunto do *powerset* (**Conjuntos Po**). Desta maneira, buscou-se mapear as regras do **Conjunto I_2E** , de modo a conhecer a sua distribuição no *ranking* (obtido pela composição de medidas de interesse, como discutido

no **Capítulo 4**) do **Conjunto Co** global. As tarefas (iv), (vi) e (vii) utilizaram valores de suporte mínimo muito baixos, de modo que a extração de regras de associação por uma abordagem convencional seria impraticável. Desta maneira, a extração de regras para as tarefas (iv) e (vi) foi interrompida, enquanto que para a tarefa (vii), não houve extração de regras.

Na **Figura 5.1**, são exibidas as posições relativas das regras obtidas com a metodologia, no **Conjunto Co**. Na representação visual, cada linha representa um conjunto contendo 1% das regras do *ranking*, de modo que as 1% melhores regras estão na primeira linha, enquanto que as 1% piores estão na base. As faixas em azul indicam a presença de regras, e a tonalidade do azul indica a quantidade de regras contida naquela faixa. O branco indica uma grande quantidade de regras, enquanto que o cinza indica ausência de regras. A escala de cores é indicada na parte superior da visualização.

Figura 5.1: Visualização das posições relativas das regras obtidas, por tarefa.

Analisando a figura, percebe-se uma distribuição relativamente uniforme das regras no caso das tarefas (ii) e (iii), enquanto que as regras resultantes das tarefas (i) e (v), foram pior ranqueadas (aparecem mais ao final da lista).

O baixo *ranking* das regras para as tarefas (i) e (v) pode ser justificado pelo fato de as regras selecionadas terem valores mais baixos para o suporte. Estes valores podem indicar que as regras envolvem itens raros, mais difíceis de serem obtidas com uma abordagem convencional e filtragem baseada no *ranking*.

A relativa uniformidade da distribuição no *ranking* das regras consideradas interessantes pelo usuário (como na **Figura 5.1 (e)**), por sua vez, pode ser explicada pelo fato

de o processo de extração com a abordagem proposta privilegiar a diversidade de itens das regras, em detrimento da obtenção pura e simples de regras com valores altos para as medidas de interesse. A abordagem busca, desta maneira, reduzir a redundância do conjunto de regras apresentado ao minerador. Assim, os conjuntos de regras obtidos com a abordagem podem ser interpretados como sumários do conhecimento contido nos dados.

Com este estudo, primeiramente, conclui-se que a metodologia é eficaz na obtenção de conjuntos de regras de associação compactos. Nas tarefas definidas para o problema apresentado, as taxas de redução do número de regras foram de 86,55% a 99,03%, em relação ao conjunto total de regras que podem ser obtidas, o *powerset*. Utilizando-se de diversas configurações de filtragem adicional, foi possível reduzir ainda mais o número de regras, variando de 88,57% a 99,51%, em relação ao *powerset*. Estas taxas de redução do número de regras são bem superiores às obtidas quando se adota uma abordagem convencional equivalente, com o uso de filtros de suporte mínimo e de confiança mínima.

Uma vez que ambos os conjuntos **I₂E** e **Co** contêm as regras de associação selecionadas pelo especialista do domínio como sendo interessantes, foi possível compará-los. Sabe-se que o conjunto de regras obtidas com a metodologia é um subconjunto do conjunto de regras obtidas com uma abordagem convencional. Verificando-se a distribuição das regras do primeiro, no segundo, percebe-se que, no caso de algumas tarefas, as regras se encontram distribuídas de maneira relativamente uniforme em todo o *ranking*, enquanto que em outras, elas se encontram distribuídas de maneira relativamente uniforme em regiões específicas. Esta distribuição relativamente uniforme ratifica a proposta de obter conjuntos de regras de associação com maior diversidade. Além disso, o fato de regras se encontrarem distribuídas em regiões revela a presença de regras com itens raros, isto é, regras com altos valores para as medidas de interesse, mas que possuem baixo suporte.

Em uma abordagem convencional, as regras mais no topo do *ranking* são certamente consultadas, enquanto que as regras mais ao final do *ranking*, apesar de interessantes, têm uma chance muito grande de serem ignoradas. Isto é devido ao fato de a exploração feita por humanos não ser constante: em situações práticas, ao deparar com conjuntos enormes de regras, os mineradores se limitam a analisar algumas regras do topo do *ranking*; quando chegam ao meio do *ranking*, se chegarem, o critério de análise pode ser diferente do critério utilizado no início. As regras da metade final do *ranking*, sendo algumas interessantes, provavelmente sequer serão consultadas.

Finalmente, considerando as regras consideradas interessantes estão localizadas na base do *ranking*, para analisá-las utilizando uma abordagem convencional, seria necessário varrer uma grande quantidade de regras para encontrá-las – o que é bastante improvável.

5.4 Avaliações da Visualização de Itemsets

5.4.1 Comparação com Outras Visualizações de Itemsets

Com o objetivo de comparar a visualização de *itemsets* da abordagem apresentada com outras soluções existentes, foram utilizadas as características utilizadas na comparação feita no levantamento bibliográfico, para apoiar a realização de tarefas de extração de regras de associação pelo minerador.

As cinco primeiras características dizem respeito à qualidade da representação dos *itemsets*, enquanto que as duas últimas correspondem a atividades recorrentes relacionadas à exploração dos *itemsets*. A comparação é mostrada na **Tabela 5.12**, em que a **Tabela 3.1** é revisitada.

Tabela 5.12: Tabela comparativa das abordagens de visualização de *itemsets*.

	Mahanti & Alhajj (2005)	Yang (2003)	Liu & Salvendy (2006)	Ertek & Demiriz (2006)	Abordagem interativa I ₂ E
Representação de <i>itemsets</i> com 1 item	Boa (primeiro nível da estrutura <i>trie</i>)	(não definido)	Boa (uma marca na interseção da linha com a coluna)	Boa (uma conexão entre um nó-item e um nó- <i>itemset</i>)	Boa (definida separadamente, em uma lista)
Representação de <i>itemsets</i> com 2+ itens	Ruim (usuário precisa percorrer a estrutura)	Boa (polilinhas longas indicam <i>itemsets</i> com mais itens)	Regular (cada relacionamento item- <i>itemset</i> é representado por uma marca)	Boa (nó adicional e arcos ligando-o aos itens; cor mapeia tamanho)	Boa (tamanho do nó indica <i>itemsets</i> com mais itens)
Representação do suporte de <i>itemsets</i>	Regular (valor é mostrado ao lado do <i>itemset</i>)	Boa (valor é mostrado na seleção; cor é indicada para cada segmento)	Boa (barras e valores são mostrados, na seleção)	Boa (tamanho do nó- <i>itemset</i> indica o valor)	Boa (valor é mostrado em texto <i>tooltip</i> ; codificação de cor é usada)
Representação do tamanho de <i>itemsets</i>	Regular (indicado pelo número de nós percorridos)	Boa (indicado pelo número de eixos tocados pela polilinha)	Regular (indicado pelo número de marcas na coluna correspondente)	Boa (indicado pela cor do nó- <i>itemset</i> correspondente)	Boa (indicado pelo tamanho do nó- <i>itemset</i> correspondente)
Escalabilidade visual	Ruim (toda a estrutura <i>trie</i> é representada)	Ruim (para cada <i>itemset</i> , uma polilinha é adicionada)	Ruim (o tamanho da matriz é dado por n <i>itemsets</i> vezes m itens)	Ruim (para cada <i>itemset</i> , um nó e arestas ligando-o aos seus itens são adicionados)	Boa (para cada <i>itemset</i> , apenas um nó é adicionado)
Seleção e exploração de um <i>itemset</i>	Ruim (o usuário precisa percorrer a estrutura)	Ruim (clicando na polilinha correspondente)	Ruim (o usuário precisa varrer toda a matriz e identificar os itens)	Boa (clicando no nó- <i>itemset</i> , as conexões indicam os seus itens)	Boa (clicando no nó- <i>itemset</i> , são mostrados seus itens na lista à parte)
Seleção, exploração e comparação de dois ou mais <i>itemsets</i>	Ruim (o usuário precisa percorrer a estrutura)	Regular (a proximidade e o ângulo dos segmentos das polilinhas indica similaridade)	Ruim (o usuário precisa varrer toda a matriz e comparar item por item)	Ruim (o usuário precisa selecionar um <i>itemset</i> por vez e comparar seus itens)	Boa (pelo posicionamento; <i>itemsets</i> similares são ligados por uma aresta)

A abordagem de (Mahanti e Alhajj, 2005) é apropriada apenas para fins educacionais, já que sua representação não favorece a busca por *itemsets*, tampouco a seleção, exploração e comparação entre dois ou mais *itemsets*. Além disso, esta abordagem não tem boa escalabilidade visual, já que ocupa uma área de visualização muito grande.

As abordagens de (Yang, 2003) e de (Liu e Salvendy, 2006) têm alguns pontos fortes. A primeira representa eficientemente *itemsets* com dois ou mais itens (mas não consegue representar *itemsets* com um item), seu suporte e tamanho. Entretanto, esta abordagem não favorece a seleção de um *itemset* (é preciso clicar sobre uma linha) e a comparação de dois ou mais *itemsets* (a única dica apresentada é a proximidade e o ângulo entre os segmentos das polilinhas, que pode indicar similaridade). A segunda, por outro lado, é mais simples, pois não requer treinamento para seu uso. Entretanto, apesar de representar suficientemente bem os *itemsets*, seu suporte e tamanho, não favorece a recuperação destas informações, especialmente do tamanho e do conteúdo (os itens propriamente ditos), já que requer uma inspeção de toda a coluna da matriz que corresponde ao *itemset*. O mesmo se aplica para a comparação entre dois ou mais *itemsets*. Além disso, a escalabilidade visual de ambas é ruim, pois as visualizações ocupam muito espaço.

A representação de (Ertek e Demiriz, 2006), por sua vez, além de representar eficientemente os *itemsets*, seu suporte e tamanho, favorece a busca por *itemsets*, destacando os itens contidos neles. Entretanto, esta representação não favorece a comparação de dois ou mais *itemsets*, tampouco os agrupa de acordo com características em comum. A solução proposta nesta tese, além de agrupar de acordo com características dos *itemsets*, tem uma escalabilidade visual melhor, já que utiliza poucos objetos para representar os *itemsets*.

5.4.2 Escolha da Métrica de Distância

É interessante analisar as características das projeções de *itemsets* geradas usando diferentes métricas de distância, em diferentes tipos de bases de dados. Para isso, foram geradas projeções para três bases de dados com características distintas – PMGRN, Orkut e Corpus Artificial, usando as duas métricas de distância disponibilizadas no Sistema I₂E. Estas projeções foram estudadas quanto ao posicionamento dos itens/*itemsets*.

A base de dados Orkut corresponde a um subconjunto dos dados obtidos do sistema Orkut, da Google, e contém informações acerca de comunidades e de seus membros, contidos em duas tabelas: Comunidades e MembrosComunidades. A primeira contém informações sobre as comunidades coletadas. A segunda, informações sobre a relação

entre membro e comunidade¹. Os atributos destas tabelas são descritos nas **Tabelas 5.13** e **5.14**, respectivamente.

Tabela 5.13: Atributos da tabela Comunidades.

Atributo	Domínio	Descrição
nome	Texto	Nome da comunidade
cmmId	Texto	ID da comunidade, gerada pelo Orkut
COM_ID	Númerico	Código da comunidade
membercount	Númerico	Quantidade de membros da comunidade
Proc_Tema	Númerico	1 = comunidade do tema (faz parte do tema pesquisado), 0 = demais

Tabela 5.14: Atributos da tabela MembrosComunidades.

Atributo	Domínio	Descrição
COM_ID	Texto	ID da comunidade
MEMBER_ID	Texto	ID do membro pertencente à comunidade

A base Corpus Artificial, por sua vez, é uma coleção de documentos criada por um gerador de corpora artificiais. O gerador, implementado com MatLab, foi obtido do LABIC (Laboratório de Inteligência Computacional) do Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP). O gerador pode criar corpora de documentos a partir de alguns parâmetros de entrada, descritos a seguir.

- Número de documentos (d): número total de documentos a serem gerados (cada um em um arquivo texto)
- Número de termos (m): número total de termos contidos nos documentos
- Número de comunidades (k): número total de comunidades/agrupamentos de documentos (todos têm o mesmo tamanho)
- Quantidade de termos tópicos (x): termos que definem uma comunidade (tem probabilidade maior de ocorrência nos documentos da comunidade)
- Probabilidade do tópico na comunidade (P): probabilidade de o termo estar em um arquivo texto pertencente a uma comunidade

De acordo com a implementação, o número de documentos (d) deve ser múltiplo do número de comunidades (k), pois o número de documentos obtidos é igual para cada

¹Por questões de privacidade, não foram utilizados atributos relacionados ao membro, com exceção do seu identificador numérico.

comunidade. Além disso, se o número de comunidades vezes o número de termos tópicos for menor do que o número total de termos ($k \cdot x < m$), então serão gerados termos comuns a todas as comunidades.

Neste estudo, foi gerado um corpus artificial com 1.000 documentos, 128 termos, sendo 16 deles tópicos, 8 comunidades (grupos de documentos) e probabilidade de um termo tópico pertencer à comunidade de 90%.

Foram utilizadas duas métricas de distância – *CityBlock* e *Euclidean* para gerar projeções, respectivamente, dos 2-*itemsets* obtidos a partir da base de dados do PMGRN fêmeas (**Figuras 5.2(a)** e **5.2(b)**), da base do Orkut (**Figuras 5.3(a)** e **5.3(b)**), e da base de dados Corpus Artificial (**Figuras 5.4(a)** e **5.4(b)**).

Figura 5.2: Visualização de 2-*itemsets*, base PMGRN fêmeas.

Figura 5.3: Visualização de 2-*itemsets*, base Orkut.

Analisando as representações visuais, percebe-se a formação dos agrupamentos de *itemsets*, utilizados na exploração. Os agrupamentos de maior destaque (os maiores)

Figura 5.4: Visualização de 2-*itemsets*, base Corpus Artificial.

são compostos pelos itens mais freqüentes nos *itemsets* apresentados, que influenciam no posicionamento dos demais agrupamentos (menores).

Para a base PMGRN, as visualizações para ambas métricas de distância (*CityBlock* e *Euclidean*) são bastante parecidas. Por outro lado, para a base Orkut, isto não acontece. Utilizando a métrica *CityBlock*, percebe-se que os agrupamentos maiores são posicionados, na sua maioria, na porção central da visualização, como pode ser visto na **Figura 5.3(a)**. Se a métrica utilizada é a *Euclidean*, estes agrupamentos são posicionados tanto no centro quanto na borda da representação visual (**Figura 5.3(b)**). Isto pode ser explicado pela natureza distinta das duas bases de dados. A primeira segmenta os itens em grupos que não podem formar *itemsets* (por exemplo, D3P_bottom_50 não pode formar um *itemset* com D3P_top_50, pois é incoerente), enquanto que a segunda não, o que limita a quantidade de possibilidades, fazendo com que o espaço de *itemsets* seja de certa forma segmentado.

Por fim, para a base Corpus, que não possui a segmentação dos itens em grupos, mas sim uma segmentação dos itens pelas comunidades, pode-se perceber que a métrica influencia no posicionamento dos grupos, porém os grupos determinados pelas comunidades (oito) se mantêm coesos.

Ao realizar tarefas de obtenção de regras de associação, um dos objetivos é encontrar regras relacionando itens menos freqüentes, evitando regras óbvias ou que caracterizam conhecimento prévio. Desta maneira, neste estudo, para a base Orkut, foram eliminados os *itemsets* que envolvem os itens mais freqüentes, resultando nas visualizações apresentadas nas **Figuras 5.5(a)** e **5.5(b)**, respectivamente, para as métricas *CityBlock* e *Euclidean*.

Figura 5.5: Visualização de 2-*itemsets*, usando diferentes métricas de distância.

Analisando as visualizações, percebe-se uma certa semelhança no posicionamento dos principais agrupamentos – aqueles que possuem *itemsets* com os itens mais frequentes. Neste ponto, já é possível perceber a formação de grupos isolados de *itemsets* ou grupos com poucas ligações – veja os agrupamentos localizados mais à esquerda nas visualizações – em ambas representações.

Selecionando um pequeno grupo de *itemsets* com baixo suporte, cujo item predominante é *Carpe_Diem*, foram obtidas as regras mostradas na **Tabela 5.15**. A partir destas correlações, o minerador pode investigar com mais profundidade os relacionamentos indiretos entre os itens, mediados pelo item *Carpe_Diem*.

Tabela 5.15: Regras com suporte baixo obtidas a partir de um agrupamento de *itemsets*.

Antecedente	Conseqüente	Suporte	Confiança	<i>Lift</i>
Adoro_vinho_tinto!	Carpe_Diem	10	0,312	4,481
Fatboy_Slim	Carpe_Diem	11	0,220	3,155
Rio_de_Janeiro	Carpe_Diem	13	0,176	2,519
Beijo	Carpe_Diem	12	0,185	2,647
Sushi_—_Japanese_food_—_??	Carpe_Diem	10	0,200	2,868
Dormir_de_Conchinha	Carpe_Diem	13	0,140	2,004

Diante do estudo analítico das visualizações obtidas, utilizando as duas métricas de distância disponíveis, podem ser feitas algumas observações. Primeiramente, o posicionamento dos *itemsets* – e conseqüentemente, dos agrupamentos – foi fortemente influenciado pela ocorrência de itens bastante frequentes (eles provocaram a formação de um grande número de *itemsets*).

Segundo, em bases em que não houve limitação do espaço de *itemsets* como o do Orkut, o posicionamento dos agrupamentos maiores – que envolvem os itens mais freqüentes e que provocam o aparecimento de um grande número de *itemsets* – foi direcionado para o centro se a métrica de distância *CityBlock* foi utilizada, enquanto que não houve um direcionamento desta natureza para a métrica *Euclidean*. No caso de bases em que houve essa limitação (como a do PMGRN), as visualizações utilizando métricas de distância diferentes tenderam a ser similares, de maneira geral.

Por fim, a eliminação de *itemsets* contendo os itens mais freqüentes permite observar com maior clareza agrupamentos isolados de *itemsets* ou com poucas ligações. Este tipo de exploração é particularmente interessante quando se busca encontrar regras envolvendo itens com suporte mais baixo, e que são dificilmente encontrados em uma exploração convencional.

5.5 Avaliação do Limiar de Filtragem

O objetivo deste estudo é investigar o efeito de utilizar diferentes limiares de filtragem, e como determinar quais seriam os limiares adequados para obter conjuntos reduzidos de regras com a abordagem proposta. Para isso, foram utilizadas bases com características bem distintas: Corpus Artificial (foi gerada uma base artificial contendo 1.000 documentos, com 128 termos tópicos e com 90% de probabilidade de estes estarem em uma comunidade), Orkut e PMGRN, descritas anteriormente. A primeira tem grupos de transações com conteúdo semelhante (muitos itens em comum), de modo que eles podem ser facilmente definidos em grupos. A segunda, por outro lado, não possui grupos facilmente distinguíveis. A última, por sua vez, é constituída por pares atributo-valor derivados de atributos que podem assumir diversos valores.

Para cada uma das bases, foram utilizados os valores de suporte típicos adotados em outros estudos: para a base de documentos, foi adotado suporte 30. Para a base Orkut, 10. Por fim, para a base do PMGRN, 292 e 583. Para cada uma das bases, foram obtidos os conjuntos de regras com percentuais arbitrários de filtragem: sem filtragem, 15%, 25% e 50% (no Sistema I₂E, pode-se indicar o valor percentual). A partir desses conjuntos, por meio de diferenças, o conjunto de regras foi particionado em quatro: conjunto **C1** (com as 50% melhores regras), conjunto **C2** (entre as 50% e as 75% melhores), conjunto **C3** (entre as 75% e as 85% melhores), e conjunto **C4** (as 15% piores).

Para analisar estes conjuntos de regras, é utilizada uma representação visual similar a um gráfico de dispersão, em que o eixo x mapeia suporte e o eixo y mapeia confiança, e composta por áreas coloridas, cada uma representando um conjunto. Uma área é

determinada pelo intervalo de valores de suporte (representado no eixo x) e de confiança (representado no eixo y) do conjunto de regras representadas. A cor indica o intervalo de regras representado: em vermelho, o conjunto **C1**; em verde, o conjunto **C2**; em azul, o conjunto **C3**, e em branco, o conjunto **C4**. As marcas coloridas em formato de cruz dentro das áreas correspondem ao valor médio do suporte e da confiança para o respectivo conjunto – uma representação esquemática é dada na **Figura 5.6**. Segundo a representação visual, uma boa maneira de escolher o limiar de corte é determinar a maior diferença entre os intervalos de suporte e de confiança de conjuntos, o que indica que as regras descartadas têm valores significativamente menores que as demais.

Figura 5.6: Representação esquemática dos conjuntos de regras.

Nas **Figuras 5.7**, **5.8** e **5.9** são mostradas as representações visuais para as bases de documentos, Orkut e PMGRN, respectivamente.

Primeiramente, para a base Corpus Artificial, um bom percentual de corte para as regras de tamanho 2 é 25% (obter as 75% melhores), pois a diferença entre as regiões das áreas é maior entre a azul (regras entre as 75% e 85% melhores) e a verde (regras entre as 50% e as 75% melhores). Para as regras de tamanho 3, as áreas branca e azul correspondem a altos valores de confiança, de modo que seria melhor indicar um percentual de corte menor que 15%, para obter mais que as 85% melhores regras.

Para a base de dados do Orkut, é adequado usar o percentual de corte 25% (obter as 75% melhores) tanto para as regras de tamanho 2 quanto para as de tamanho 3, pois as áreas em azul e em branco indicam conjuntos de regras com baixos valores de suporte e de confiança. Entretanto, pode-se pensar em aumentar o percentual para 50% para as regras de tamanho 2 devido ao baixo valor de confiança, e para as regras de tamanho 3 devido ao baixo valor de suporte.

(a) Conjuntos de regras de tamanho 2 (b) Conjuntos de regras de tamanho 3

Figura 5.7: Análise de limiar de filtragem para a base Corpus Artificial.

(a) Conjuntos de regras de tamanho 2 (b) Conjuntos de regras de tamanho 3

Figura 5.8: Análise de limiar de filtragem para a base Orkut.

Por fim, fez-se a mesma análise na base do PMGRN. Primeiramente, para os valores de suporte 292 (**Figura 5.9(a)**) e (**Figura 5.9(b)**) e regras de tamanho 2, o valor do percentual de corte adequado é 75%, pois as regiões em azul e em branco apresentam baixos valores de suporte e de confiança. Para suporte 583 e regras de tamanho 3, o valor mais adequado seria 15%, já que a área em azul cobre uma área de alto valor de confiança.

Observando os resultados com essas três bases de dados, é possível delinear diretrizes gerais para a determinação do limiar de filtragem. Limiares altos, por exemplo, de 50% a 75%, podem reduzir sensivelmente a quantidade de regras, mas também eliminar regras com valores razoáveis de medidas de interesse, o que não vai ocorrer se o filtro for conservador (de 85% a sem filtro). Uma boa alternativa é configurá-lo dentro do intervalo 75%-85%, pois as regras com baixos valores são eliminadas (geralmente, ficam entre as 15% piores), enquanto que regras possivelmente interessantes são preservadas.

(a) Conjuntos de regras de tamanho 2 para suporte 292 (b) Conjuntos de regras de tamanho 2 para suporte 583 (c) Conjuntos de regras de tamanho 3 para suporte 583

Figura 5.9: Análise de limiar de filtragem para a base PMGRN.

5.6 Considerações Finais

Neste capítulo, foram descritas diversas avaliações da metodologia introduzida no **Capítulo 4** e o seu uso. De uma maneira geral, elas buscaram: (1) estudar o comportamento das regras extraídas com a metodologia sob diversos aspectos, em relação ao conjunto de regras obtido por uma abordagem convencional; e (2) analisar o processo, com relação a alguns parâmetros de execução.

Diante dos resultados, observamos que a metodologia proposta se mostrou mais adequada à execução de certas tarefas de extração de regras de associação do que abordagens convencionais, por prover recursos capazes de evitar que execuções inteiras sejam feitas antes que uma análise dos resultados seja possível, por gerar conjuntos de regras mais compactos, por preservar a cobertura de *itemsets*, por favorecer a reformulação de tarefas ou a formulação de novas tarefas, e por prover meios para comparação visual de regras, aumentando o poder de análise.

No próximo capítulo, são apresentados estudos de caso em situações reais, com usuários reais utilizando a abordagem.

Avaliação da Abordagem: Estudos de Caso com Usuário

6.1 Considerações Iniciais

Foram realizados estudos de caso com usuários reais com o objetivo de verificar o potencial de uso da abordagem proposta. Primeiramente, foram conduzidos dois estudos de caso com especialista de um domínio (base PMGRN). Depois, são relatados dois estudos de caso de obtenção de regras de associação com a abordagem interativa em um cenário exploratório (com a base Orkut), conduzidos por usuários finais.

No primeiro estudo de caso, o usuário do sistema é um especialista do domínio (PMGRN), pós-graduando em Genética pela Faculdade de Medicina de Ribeirão Preto (FMRP) da Universidade de São Paulo (USP), realizou diversas explorações (cerca de quatro) sem treinamento prévio, mas com acompanhamento para que dúvidas pontuais fossem solucionadas. As explorações duraram efetivamente cerca de meia hora, para a realização das tarefas. No primeiro encontro, o acompanhamento foi mais próximo: foram dadas explicações sobre como o sistema funciona com exemplos, enquanto que nos demais, o usuário apenas solucionou dúvidas.

No segundo estudo de caso, os usuários do sistemas foram/são pós-graduandos em Ciências de Computação pelo Instituto de Ciências Matemáticas e de Computação (ICMC)

da Universidade de São Paulo (USP). O usuário da primeira parte do estudo é especialista em inteligência artificial/mineração de dados. Ele participou da segunda parte do estudo juntamente com outro usuário, que é especialista em mineração de dados e em visualização de informação. Em cada parte do estudo, foi feito apenas um encontro com cada usuário.

Na primeira parte do estudo, o usuário não teve treinamento, porém teve acompanhamento próximo, sendo dadas explicações sobre como o sistema funciona com exemplos e dúvidas foram solucionadas. Nesta ocasião, o usuário utilizou o sistema por pouco tempo (menos de meia-hora) para explorar os *itemsets* freqüentes, gastando mais tempo na exploração e análise das regras (cerca de duas horas). O segundo usuário também não teve treinamento, tendo acompanhamento para solução de dúvidas. Ele, por sua vez, gastou tempo considerável na exploração de *itemsets* freqüentes (cerca de meia-hora), gerando uma quantidade pequena de regras, o que fez com que a exploração do conjunto de regras fosse simplificada.

6.2 Estudos de caso com Especialista do Domínio

6.2.1 Primeiro estudo de caso – PMGRN fêmeas

O objetivo deste estudo foi conduzir todo um processo de descoberta de regras de associação com a abordagem proposta, a partir de uma base de dados real, visando explorar os seus recursos e verificar a sua utilidade. Neste estudo, foi utilizado um subconjunto da base de dados do PMGRN – 17 atributos (*categoria*, *variedade*, *USA Classe*, *USA Ano*, *USA Mes*, *NFA*, *regiao*, *MP120*, *DP120*, *DP365*, *DPE365*, *DPAC*, *DSTAY* e *D3P*), formando 67 pares atributo-valor, contendo informações sobre fêmeas vendidas.

Do ponto vista do especialista do domínio, a descoberta de regras de associação nesta base tinha como objetivos comerciais: aumentar a cartela de clientes do programa, atraindo novos produtores por meio de campanhas de *marketing* que mostrem alternativas para possíveis vantagens competitivas no mercado; e fornecer informações relevantes a clientes do programa na escolha de touros e matrizes (fêmeas) para acasalamento e inseminação.

Inicialmente, o especialista formulou quatro tarefas-alvo:

1. Identificar um perfil dos animais comercializados por ano.
2. Identificar um perfil dos animais comercializados por região.
3. Identificar características específicas de animais da classe “novilha” comercializados.

4. Identificar características específicas de animais da categoria “PO” comercializados.

Após carregar a base de dados no sistema I₂E e determinar um valor de suporte arbitrário 562 (~15%), os usuários (o minerador de dados e o especialista do domínio) observaram a visualização dos *itemsets* de tamanho 2 (ver **Figura 6.1**), gerada utilizando a métrica *CityBlock*, na qual são mostrados 153 *itemsets* extraídos, separados em 7 agrupamentos (indicados pela cor dos seus *itemsets*). O tamanho dos agrupamentos de *itemsets* é indicado no histograma, apresentado no canto inferior esquerdo da janela da visualização.

Depois de explorar um pouco a visualização, 19 itens distintos (de um total de 37) e 25 *itemsets* distintos (de um total de 153) foram selecionados. Durante a exploração, eles identificaram os itens “D3P_NA” e “DSTAY_NA”, que descrevem valores ausentes na base de dados: valores não informados para D3P e DSTAY. Eles optaram por remover do processo de extração todos os *itemsets* contendo estes itens (9 *itemsets* no total).

Figura 6.1: Grupos de *itemsets* extraídos da base PMGRN fêmeas.

Ao final, foram obtidos sete grupos: um grupo com clara predominância do item Var-P (todos os 31 *itemsets* do grupo), um grupo com predominância do item Classe-Novilha (todos os 22 *itemsets*), um com DPAC_bottom_50 (todos os 10 *itemsets*), um com predominância de Cat-PO (16 *itemsets* de um total de 26), um com predominância de Reg-Centro-Oeste (todos os 21 *itemsets*), um com Fazenda-H (9 de um total de 14), e um

grupo sem predominância de nenhum item (com 19 *itemsets*). Para a redução adicional da cardinalidade do conjunto de *itemsets* de tamanho 2, foi utilizado um limiar de 85% (filtragem adicional de 15% piores regras segundo ranqueamento com múltiplas medidas de interesse), resultando em 122 regras de tamanho 2 no final desta etapa.

O processo foi repetido com *itemsets* de tamanho 3 (95 no total), 4 (39) e 5 (17). Nestas etapas, todos os *itemsets* foram considerados relevantes, e nenhuma remoção adicional foi feita. Os *itemsets* de tamanho 3 foram reorganizados em quatro grupos: um grupo com predominância de Cat-PO (todos os 5 *itemsets*), um grupo com predominância de Var-P (49 de um total de 52), um com predominância de DP365_bottom_50 (14 de um total de 25) e finalmente um com predominância de Reg-Centro-Oeste (todos os 13) *itemsets*. Mais uma vez, o limiar 85% foi utilizado para o filtro, gerando 80 regras nesta etapa. Os *itemsets* de tamanho 4 foram reorganizados em 6 grupos, sendo um deles sem nenhum item predominante – nos outros, os *itemsets* eram apenas combinações diferentes de alguns itens – e 33 regras foram obtidas. Todos os *itemsets* de tamanho 5 (17) eram combinações diferentes de alguns itens, sendo que nenhum grupo foi formado, e 17 regras foram geradas. O mesmo ocorreu para os 4 *itemsets* de tamanho 6, e 4 regras foram geradas.

Ao final do processo, foi obtido um conjunto global com 256 regras. Com uma extração de regras convencional, usando o mesmo nível de suporte mínimo e sem filtragem por medidas de interesse o conjunto global gerado teria 2.314 regras. Por outro lado, se fosse aplicada uma filtragem *a posteriori* baseada simplesmente em um ranqueamento, o conjunto de regras resultante teria diversidade sensivelmente menor.

O conjunto global de regras foi explorado com o módulo RulEx do Sistema I₂E. Para executar as tarefas formuladas, os mineradores decidiram focalizar inicialmente nas regras de tamanho 2. Primeiramente, eles selecionaram regras com itens relacionados ao ano, listadas na **Tabela 6.1**. Desta lista, 8 regras foram subjetivamente consideradas interessantes pelo especialista, indicadas por um asterisco (*) na tabela.

Tabela 6.1: Regras relacionadas ao ano de comercialização das fêmeas.

	Antecedente	Consequente	Suporte	Confiança	Lift
*	U2004	DPAC_bottom_50	749	0,683	1,291
	U2004	Var-P	913	0,833	1,023
	U2004	Classe-Novilha	652	0,595	1,039
*	U2004	Reg-Sudeste	565	0,516	1,130
*	U2004	MP120_bottom_50	608	0,555	1,362
*	U2004	DP365_bottom_50	638	0,582	1,254
*	U2004	DP120_bottom_50	621	0,567	1,297
	U2005	Var-P	784	0,860	1,056
*	U2005	Cat-PO	825	0,905	1,813
*	U2005	D3P_bottom_50	592	0,649	1,383
*	U2006	Mês-Fevereiro	650	0,835	3,518
	U2006	Var-P	598	0,769	0,944

Eles revelam que a maioria das fêmeas vendidas no ano de 2004 (atributo U2004) eram provenientes da região Sudeste do país e tinham muitas características entre as 50% piores (bottom_50), tais como DPAC, MP120, DP120 e DP365. Em 2005, a maior parte das fêmeas vendidas era da categoria PO e tinha valor bottom_50 para D3P. Por fim, em 2006, a maior parte das fêmeas foi vendida no mês de fevereiro. Este reduzido conjunto de regras pode guiar outras explorações do espaço de regras. Por exemplo, ao inspecionar as regras com o item Var-P, que se referem a uma particular variedade de animais comercializados, pode-se observar uma tendência decrescente no período de 2004 a 2006. Em 2004, 913 animais desta variedade foram comercializados, contra 784 em 2005 e 598 em 2006. Outra observação interessante é a diminuição do número de regras envolvendo DEPs com valores bottom_50. Para o suporte adotado, foram extraídas quatro regras para o ano 2004 com valores na faixa bottom_50: DPAC, MP120, DP365 e DP120. Em 2005, observou-se apenas uma ocorrência de DEP dentro desta faixa (D3P), enquanto que nenhuma ocorrência de DEPs com valor baixo envolvendo o ano 2006 foi observada.

Para a segunda tarefa-alvo, os mineradores focalizaram as regras envolvendo a região: aquelas consideradas interessantes são mostradas na **Tabela 6.2**. Observou-se que a maioria das fêmeas da região Sudeste eram da categoria PO, com valores bottom_50 para a maioria das DEPs (DPE365, MP120, DP365, DP120 e DPAC). Por outro lado, na região Centro-Oeste, a maioria das fêmeas era proveniente da fazenda H. Além disso, havia uma predominância de fêmeas da categoria LA, de variedade M, com valores top_25 para a maioria das DEPs (DP120, DP365, DPAC, DSTAY, DPE365, DP120 e D3P), com valor top_50 apenas para DP120 e bottom_50 apenas para DSTAY. Isto significa que, em geral, as fêmeas vendidas provenientes desta região tinham boas características genéticas, o que não ocorreu com as provenientes da região Sudeste, como já foi observado.

Na terceira tarefa, foram consideradas fêmeas da classe Novilha. As regras selecionadas são apresentadas na **Tabela 6.3**. Analisando essas regras, observa-se que a maioria das novilhas comercializadas eram provenientes da região Centro-Oeste, a maior parte delas da fazenda H. Além disso, estes animais tinham valores top_25 para a maior parte das DEPs analisadas: DP120, DP365, DPE365 e D3P.

Por fim, as regras interessantes envolvendo fêmeas de categoria PO são mostradas na **Tabela 6.4**. Observa-se que a maior parte das fêmeas de variedade P eram de categoria PO e tinham valores no intervalo top_25 para DP365, DPAC, DSTAY e MP120, valores no intervalo top_50 para MP120, e valores no intervalo bottom_50 para D3P e DPE365. Além disso, todas as fêmeas comercializadas provenientes da fazenda F eram da categoria PO. Outra observação importante é que há uma maior quantidade de ocorrências de MP120_top_25 do que de MP120_top_50.

Tabela 6.2: Regras relacionadas à região de origem das fêmeas comercializadas.

Antecedente	Conseqüente	Suporte	Confiança	<i>Lift</i>
Reg-Sudeste	Cat-PO	1.111	0,652	1,306
Reg-Sudeste	DPE365_bottom_50	978	0,574	1,310
Reg-Sudeste	MP120_bottom_50	841	0,493	1,211
Reg-Sudeste	DP365_bottom_50	996	0,584	1,259
Reg-Sudeste	DP120_bottom_50	931	0,546	1,249
Reg-Sudeste	DPAC_bottom_50	998	0,585	1,105
Reg-Centro-Oeste	Fazenda-F	577	0,304	1,970
Reg-Centro-Oeste	Fazenda-H	1.319	0,695	1,970
Reg-Centro-Oeste	Cat-LA	797	0,420	1,448
Reg-Centro-Oeste	Var-M	691	0,364	1,959
Reg-Centro-Oeste	DP120_top_25	710	0,374	1,259
Reg-Centro-Oeste	DP365_top_25	756	0,398	1,304
Reg-Centro-Oeste	DPAC_top_25	610	0,321	1,155
Reg-Centro-Oeste	DSTAY_top_25	748	0,394	1,217
Reg-Centro-Oeste	DPE365_top_25	898	0,473	1,351
Reg-Centro-Oeste	MP120_top_25	806	0,425	1,306
Reg-Centro-Oeste	DP120_top_50	581	0,306	1,151
Reg-Centro-Oeste	D3P_top_25	704	0,371	1,421
Reg-Centro-Oeste	DSTAY_bottom_50	715	0,377	1,171

Tabela 6.3: Regras relacionadas à classe Novilha para as fêmeas comercializadas.

Antecedente	Conseqüente	Suporte	Confiança	<i>Lift</i>
DP120_top_25	Classe-Novilha	772	0,695	1,214
DP365_top_25	Classe-Novilha	775	0,679	1,186
DPE365_top_25	Classe-Novilha	868	0,663	1,159
Fazenda-H	Classe-Novilha	838	0,635	1,110
Reg-Centro-Oeste	Classe-Novilha	1.145	0,603	1,054
D3P_top_25	Classe-Novilha	618	0,633	1,106

Tabela 6.4: Regras relacionadas à categoria PO para as fêmeas comercializadas.

Antecedente	Conseqüente	Suporte	Confiança	<i>Lift</i>
Var-P	Cat-PO	1.598	0,525	1,052
Fazenda-F	Cat-PO	577	1,000	2,004
DP365_top_25	Cat-PO	583	0,511	1,023
DPAC_top_25	Cat-PO	709	0,682	1,366
DSTAY_top_25	Cat-PO	903	0,746	1,494
MP120_top_25	Cat-PO	662	0,544	1,091
MP120_top_50	Cat-PO	582	0,582	1,166
DP120_top_50	Cat-PO	569	0,572	1,147
D3P_bottom_50	Cat-PO	1.187	0,676	1,355
DPE365_bottom_50	Cat-PO	841	0,514	1,029

Como o especialista do domínio havia formulado quatro tarefas-alvo, o processo pôde ser direcionado para obter as regras envolvendo os itens de interesse. Inicialmente, pode-se pensar que é possível resolver estes tipos de tarefa com uma abordagem online ou especificando restrições. De fato, isso seria possível para alguns itens como Classe-Novilha, mas não para outros como itens envolvendo ano ou região, que demandariam uma alta carga cognitiva por parte do usuário, que deveria ser capaz de formular todas as restrições a serem especificadas. Utilizando uma abordagem pró-ativa como a apresentada, o usuário

pode explorar todas as possibilidades de associação entre os itens, o que favorece a inferência de novas associações não antecipadas como tarefas-alvo. Esta interação enriquece a atividade, e auxilia indiretamente o usuário a compreender melhor a base de dados.

6.2.2 Segundo estudo de caso – PMGRN machos

O objetivo deste estudo é similar ao anterior, mas foi particularmente observada a possibilidade de reformulação, durante o processo de extração de regras, dos objetivos traçados pelo especialista. Este estudo foi focado em um subconjunto da base de dados PMGRN – 15 atributos (*NFA*, *Regiao*, *USA Classe*, *USA Ano*, *USA Mes*, *categoria*, *variedade*, *MGT*, *MP120*, *DP365*, *DPE365*, *DIPP*, *DSTAY*, *DAOL*, *DEG*), contendo informações sobre machos comercializados.

Primeiramente, o especialista do domínio determinou quatro tarefas a cumprir:

1. Identificar especializações das fazendas na comercialização de touros quanto ao perfil genético.
2. Identificar o perfil de touros comercializados das categorias CL e LA.
3. Verificar mudança no perfil genético dos animais comercializados com o passar dos anos.
4. Verificar o período do ano com maior comercialização dos touros, para cada fazenda.

O especialista do domínio explorou a base de dados utilizando o Sistema I₂E com o auxílio do minerador.

Na **Tarefa 1**, o especialista buscava identificar particularidades de fazendas na comercialização de touros quanto ao perfil genético. Ele decidiu selecionar regras que envolvessem itens contendo uma fazenda e um valor de DEP específicos. Inicialmente, o valor do suporte mínimo foi ajustado para 583 ($\sim 10,49\%$). A exploração indicou três fazendas (B, F e H) – na **Figura 6.2**, é mostrada a seleção de *itemsets* envolvendo a fazenda H. Entretanto, apenas para as fazendas F e H foram geradas regras interessantes. A fim de obter regras interessantes para a fazenda B, o suporte mínimo foi reduzido para 292 ($\sim 5,25\%$). Com este novo valor foram obtidas regras para as fazendas A e B, mas apenas as que envolviam a fazenda B foram consideradas interessantes.

A política de redução de regras utilizada foi *um-por-itemset*, e o agrupamento *automático* (com o algoritmo *k*-médias) foi executado utilizando a métrica *CityBlock*. Esta configuração foi usada para todas as tarefas. A partir daí, um conjunto reduzido de regras,

Figura 6.2: Visualização de 2-itemsets da base PMGRN machos.

mostrado na **Tabela 6.5**, foi explorado com o módulo RuleEx. O especialista decidiu focar nas regras com valor de $lift \geq 1,1$. Analisando estas regras, foi possível identificar perfis genéticos distintos dos touros, originados de diferentes fazendas. Enquanto a fazenda F é especializada na venda de touros de alta qualidade para DSTAY e DPAC, a fazenda H é especializada em touros de alta qualidade para DPE365, DP120, DP365 e D3P. A fazenda B, por sua vez, tem uma especialização distinta: touros de alta qualidade para MP120. Este tipo de informação é importante para que compradores saibam quais fazendas têm mais probabilidade de ter touros de uma determinada característica. Desta maneira, eles podem direcionar o processo de busca para aquisição de novos animais. Esta informação também seria útil em campanhas de *marketing* para atrair mais criadores para o programa.

Enquanto executava as tarefas, o especialista sugeriu que seria interessante buscar perfis mais específicos dos touros comercializados em cada fazenda. Assim, ele gerou regras que envolvessem três itens (dois para valores de DEP e um para fazenda). O suporte mínimo foi mantido em 583 ($\sim 10,49\%$), assim como o $lift$ mínimo, que permaneceu em 1,1. Com estes valores, foram geradas regras para as fazendas B, F e H, mas apenas regras envolvendo as fazendas F e H foram consideradas interessantes. Estas regras são mostradas na **Tabela 6.6**.

Tabela 6.5: Algumas das regras envolvendo as fazendas F, H e B.

N.	Antecedente	Consequente	Suporte	Confiança	Lift
1.1	Fazenda-F	DSTAY_top_25	1.273	0,79	1,75
1.2	Fazenda-F	DPAC_top_25	901	0,56	1,39
1.3	Fazenda-F	DPE365_bottom_50	639	0,39	1,03
1.4	Fazenda-F	DPE365_top_25	703	0,43	1,01
1.5	Fazenda-F	D3P_bottom_50	1.193	0,74	1,38
1.6	D3P_bottom_50	Fazenda-F	1.193	0,40	1,38
1.7	Fazenda-H	DPE365_top_25	1.075	0,51	1,19
1.8	Fazenda-H	DP365_top_25	1.020	0,49	1,97
1.9	Fazenda-H	DP120_top_25	951	0,45	1,12
1.10	D3P_top_25	Fazenda-H	1.025	0,61	1,61
1.11	DP120_top_50	Fazenda-H	670	0,41	1,10
1.12	Fazenda-H	DPAC_bottom_50	1.019	0,49	1,27
1.13	DSTAY_bottom_50	Fazenda-H	1.054	0,65	1,72
1.14	Fazenda-B	MP120_top_25	441	0,50	1,14
1.15	Fazenda-B	DP120_bottom_50	332	0,38	1,24
1.16	Fazenda-B	DPE365_bottom_50	390	0,44	1,16
1.17	Fazenda-B	DP120_top_50	293	0,33	1,14

Tabela 6.6: Regras envolvendo uma fazenda (F, H), e dois valores de DEP.

N.	Antecedente	Consequente	Suporte	Confiança	Lift
2.1	DSTAY_top_25,D3P_bottom_50	Fazenda-F	968	0,62	2,14
2.2	D3P_bottom_50,DPAC_top_25	Fazenda-F	652	0,52	1,79
2.3	DSTAY_top_25,DPAC_top_25	Fazenda-F	815	0,52	1,77
2.4	DSTAY_top_25,MP120_top_25	Fazenda-F	615	0,49	1,69
2.5	Fazenda-F	DSTAY_top_25,DPAC_top_25	815	0,50	1,77
2.6	Fazenda-F	DSTAY_top_25,MP120_top_25	615	0,38	1,69
2.7	DPAC_bottom_50,DSTAY_bottom_50	Fazenda-H	699	0,65	1,73
2.8	DP120_top_25,DP365_top_25	Fazenda-H	766	0,45	1,19
2.9	DPE365_top_25,D3P_top_25	Fazenda-H	661	0,64	1,70
2.10	DPE365_top_25,DP365_top_25	Fazenda-H	687	0,47	1,25
2.11	Fazenda-H	DPAC_bottom_50,DSTAY_bottom_50	699	0,33	1,73
2.12	Fazenda-H	DP120_top_25,DP365_top_25	766	0,36	1,19
2.13	Fazenda-H	DPE365_top_25,D3P_top_25	661	0,31	1,70
2.14	Fazenda-H	DPE365_top_25,DP365_top_25	687	0,33	1,25

As regras com altos valores de suporte revelam fazendas importantes como pontos de venda de touros com determinadas características genéticas. Por exemplo, a regra 2.1 (com valor de suporte 968) e as regras 2.3 e 2.5 (com valor de suporte 815). Além disso, regras com altos valores de confiança revelam as fazendas mais importantes com determinadas características genéticas (regras 2.1 a 2.4, e 2.7 a 2.10), ou especializações de cada fazenda (regras 2.5, 2.6 e 2.11 a 2.14). O especialista descobriu que a fazenda F é uma importante fonte de touros de alta qualidade para DSTAY e de baixa qualidade para D3P, com suporte 968. 62% das transações envolvendo touros com estas características foram provenientes desta fazenda. Adicionalmente, a fazenda F tem um importante papel na venda de touros: de alta qualidade para DPAC e de baixa qualidade para D3P (652 vendas, que corresponde a 52% das vendas de todas as 10 fazendas); de alta qualidade para DSTAY e DPAC (815, 52% de todas as fazendas); e de alta qualidade para DSTAY e MP120 (615, 38% de todas as fazendas). Analisando as regras 2.5 e 2.6, pudemos verificar

ainda que a fazenda F é um importante vendedor de touros de alta qualidade para DSTAY e DPAC, e de alta qualidade para DSTAY e MP120: a fazenda foca no comércio de animais com este perfil, correspondendo a 50% e a 38% do seu total de vendas, respectivamente.

Em contraposição à fazenda F, o foco das vendas da fazenda H foi em touros de baixa qualidade para DPAC e DSTAY. Entretanto, a fazenda H teve um importante papel no comércio de touros de alta qualidade para DP120 e DP365, para DPE365 e D3P, e para DPE365 e DP365, correspondendo respectivamente a 45%, 64% e 47% das vendas de todas as fazendas. Considerando apenas as vendas da fazenda H, verificou-se um foco nos perfis mencionados: touros de alta qualidade para DP120 e DP365, DPE365 e D3P, correspondendo respectivamente a 33%, 31% e 33% das suas vendas.

Aprofundando a análise, o especialista comparou os valores das medidas de interesse para a presença ou não de alguns itens para DEP nas regras. Comparando-se as regras $\text{Herd-F} \rightarrow \text{DSTAY_top_25}$ (suporte=1.273, confiança=0,79, *lift*=1,75), $\text{Herd-F} \rightarrow \text{DSTAY_top_25, DPAC_top_25}$ (815, 0,50, 1,77) e $\text{Herd-F} \rightarrow \text{DPAC_top_25}$ (901, 0,56, 1,39), o especialista observou que, adicionando-se DPAC_top_25 à primeira regra (resultando na segunda regra), os valores das medidas de interesse diminuíram significativamente, o que indica que a adição enfraquece a regra de maneira severa. Esta comparação é mostrada na **Figura 6.3**. Note que a regra $\text{Herd-F} \rightarrow \text{DSTAY_top_25}$, representada pela polilinha em vermelho, tem claramente valores maiores para as medidas de interesse que a regra $\text{Herd-F} \rightarrow \text{DSTAY_top_25, DPAC_top_25}$, mostrada em verde. Por outro lado, comparando esta observação com a possível remoção de DSTAY (resultando na terceira regra), não se nota um aumento significativo dos valores das medidas de interesse, o que indica que para a fazenda F, DSTAY_top_25 é uma característica mais forte que DPAC_top_25. Esta comparação é mostrada na **Figura 6.4**. Note que os valores para as regras são próximos (as polilinhas tocam os eixos em posições próximas).

Uma porcentagem significativa dos touros vendidos eram das categorias LA (~19,8%) e CL (~22,9%). Na **Tarefa 2**, o especialista procurou traçar um perfil genético dos touros dessas categorias. Assim, ele procurou por regras contendo itens representando uma categoria, uma fazenda e um valor de DEP específicos. Nesta tarefa, o suporte mínimo foi ajustado em 195 (~3,51%), pois nenhuma regra interessante foi obtida quando esse valor foi inicialmente ajustado em 583 (~10,49%), ou em 292 (~5,25%). Ao final, para o valor de suporte definido, apenas regras para a fazenda H foram obtidas. Estas regras, mostradas na **Tabela 6.7**, revelam que os touros destas categorias eram de alta qualidade para as DEPs, apesar de haver lacunas na árvore genealógica. Os touros de categoria CL tinham alta qualidade para D3P, MP120, DSTAY e DPAC, e média qualidade para DSTAY. Já os touros de categoria LA eram de alta qualidade para DSTAY, MP120, DPAC, D3P

Figura 6.3: Comparação de regras usando o módulo RulEx.

e DP365. Isto indica um interesse no comércio de touros destas categorias (apenas nos animais que apresentam boas características genéticas). Esta informação pode ser útil em campanhas de *marketing* do programa direcionadas para criadores cuja maior parte do rebanho é formada por animais destas categorias.

Tabela 6.7: Regras envolvendo uma categoria, um valor de DEP e uma fazenda.

N.	Antecedente	Consequente	Suporte	Confiança	Lift
3.1	Cat-CL,D3P_top_25	Fazenda-H	489	0,99	2,61
3.2	Cat-CL,MP120_top_25	Fazenda-H	567	0,98	2,59
3.3	Cat-CL,DSTAY_top_25	Fazenda-H	227	0,98	2,59
3.4	Cat-CL,DPAC_top_25	Fazenda-H	296	0,97	2,57
3.5	Cat-CL,DSTAY_top_50	Fazenda-H	245	0,96	2,55
3.6	DSTAY_top_25,Cat-LA	Fazenda-H	295	0,99	2,62
3.7	MP120_top_25,Cat-LA	Fazenda-H	455	0,99	2,62
3.8	DPAC_top_25,Cat-LA	Fazenda-H	322	0,97	2,57
3.9	D3P_top_25,Cat-LA	Fazenda-H	487	0,97	2,57
3.10	DP365_top_25,Cat-LA	Fazenda-H	541	0,96	2,55

Figura 6.4: Comparação de regras usando o módulo RulEx (continuação).

Para ter uma idéia geral da evolução dos valores para as DEPs ao longo dos anos, o especialista necessitava extrair regras que envolvessem um item temporal (o ano) e um item com determinado valor de DEP (**Tarefa 3**). Esta informação é relevante para analisar a evolução da qualidade dos rebanhos, a expectativa dos compradores ou ambas. O valor de suporte mínimo foi ajustado em 195 ($\sim 3,51\%$). Com este valor, apenas regras contendo itens que mapeiam anos de 2002 a 2005 foram geradas. As 21 regras com valores mais altos para confiança e *lift* selecionadas como sendo interessantes são mostradas na **Tabela 6.8**. Analisando cada DEP, não se identificou uma tendência clara de melhorar a qualidade genética dos touros, o que indica que a qualidade dos touros não é uma preocupação para os criadores. Entretanto, esta informação pode ser utilizada para motivá-los a melhorar a qualidade genética dos seus touros, que podem ser vendidos a preços mais elevados.

Baseado na análise da evolução dos valores das DEPs dos touros ao longo dos anos, o especialista necessitava extrair regras envolvendo o período do ano e os vendedores (**Tarefa 4**). Estas regras são particularmente interessantes, porque podem revelar sazonalidade nas

Tabela 6.8: Regras envolvendo ano e valor de DEP.

N.	Antecedente	Conseqüente	Suporte	Confiança	Lift
4.1	U2002	DP365_bottom_50	321	0,50	1,52
4.2	U2002	D3P_bottom_50	430	0,67	1,25
4.3	U2002	DP120_bottom_50	308	0,48	1,57
4.4	U2002	DSTAY_top_25	372	0,58	1,28
4.5	U2003	D3P_top_25	619	0,47	1,54
4.6	U2003	DP365_top_25	664	0,50	1,24
4.7	U2003	DPE365_top_25	680	0,52	1,20
4.8	U2003	DSTAY_bottom_50	545	0,41	1,42
4.9	U2003	DP120_top_25	620	0,47	1,16
4.10	U2003	DP120_top_50	442	0,33	1,15
4.11	U2004	DPAC_bottom_50	700	0,52	1,37
4.12	U2004	MP120_bottom_50	497	0,37	1,33
4.13	U2004	DP120_bottom_50	503	0,38	1,23
4.14	U2004	DP365_bottom_50	516	0,39	1,18
4.15	U2004	DSTAY_bottom_50	455	0,34	1,17
4.16	U2005	DSTAY_top_25	508	0,65	1,44
4.17	U2005	DPAC_top_25	425	0,54	1,36
4.18	U2005	DP365_top_25	424	0,54	1,34
4.19	U2005	DPE365_top_25	435	0,56	1,30
4.20	U2005	DP120_top_25	419	0,54	1,33
4.21	U2005	D3P_bottom_50	486	0,62	1,17

vendas. Inicialmente, o suporte mínimo foi ajustado em 292 ($\sim 5,25\%$), mas poucas regras envolvendo os atributos de interesse foram obtidas. Assim, o valor do suporte mínimo foi ajustado novamente, desta vez em 98 ($\sim 1,76\%$), e regras com $lift \geq 1,1$ foram selecionadas. Ao final da tarefa, foi obtido um conjunto de regras envolvendo cinco fazendas em seis meses diferentes, mostrado na **Tabela 6.9**.

Tabela 6.9: Regras envolvendo atributos de mês e de fazenda.

N.	Antecedente	Conseqüente	Suporte	Confiança	Lift
5.1	Mês-Maio	Fazenda-B	231	0,47	2,94
5.2	Mês-Julho	Fazenda-B	305	0,43	2,72
5.3	Mês-Outubro	Fazenda-B	99	0,17	1,05
5.4	Fazenda-B	Mês-Setembro	120	0,14	0,63
5.5	Fazenda-C	Mês-Novembro	141	0,78	1,76
5.6	Fazenda-E	Mês-Outubro	270	0,94	8,78
5.7	Mês-Agosto	Fazenda-F	380	0,77	2,63
5.8	Mês-Junho	Fazenda-F	158	0,66	2,25
5.9	Mês-Abril	Fazenda-F	199	0,62	2,13
5.10	Mês-Janeiro	Fazenda-F	275	0,26	0,89
5.11	Mês-Outubro	Fazenda-F	123	0,21	0,71
5.12	Mês-Julho	Fazenda-F	99	0,14	0,48
5.13	Mês-Setembro	Fazenda-F	152	0,13	0,43
5.14	Mês-Setembro	Fazenda-H	926	0,76	2,02
5.15	Mês-Janeiro	Fazenda-H	703	0,66	1,75
5.16	Mês-Julho	Fazenda-H	216	0,30	0,81
5.17	Mês-Maio	Fazenda-H	113	0,23	0,60

Analisando as regras, notou-se que cada fazenda tem o seu pico de comércio de touros, provavelmente, devido a leilões que promoveram. Entre os maiores criadores, a fazenda H tem picos em janeiro (66% de todas as vendas neste mês) e setembro (76%), enquanto que a fazenda F tem picos em abril (62%), junho (66%) e agosto (77%). A fazenda B, um criador

de médio porte, tem picos em maio (47%) e julho (43%). Entre os pequenos criadores, a fazenda E concentra a maior parte das suas vendas (94%) em outubro, enquanto que a fazenda C concentra as suas vendas em novembro (78%). Estes meses são considerados bons para o comércio de touros, para cada criador.

O aquecimento do mercado pode ser observado de abril a novembro, devido a dois fatores principais. Primeiramente, na maior parte das fazendas da região Centro-Oeste do país, ocorre acasalamento sazonal de novembro a janeiro, o que faz com que os compradores antecipem as aquisições (em dezembro, poucos touros são comercializados). O segundo fator é a realização das principais feiras e exposições agropecuárias em abril, dando início ao período de comercialização de touros. A fazenda H é uma exceção a esta regra: o seu pico de vendas é em janeiro.

Estas informações podem ser usadas por criadores para direcionar campanhas de *marketing* a seus clientes. Além disso, muitos criadores não têm informação precisa sobre as suas próprias vendas (de touros).

Ao longo do processo, o especialista mudou diversas vezes o valor do suporte mínimo, com o objetivo de completar as tarefas definidas previamente, mas apenas uma execução completa foi feita. Em uma abordagem convencional, seria preciso repetir o processo inúmeras vezes, até encontrar algum resultado aceitável.

Adicionalmente, enquanto executava a **Tarefa 1**, o especialista reformulou uma tarefa, decidindo investigar mais a fundo as particularidades de vendas das fazendas. Outra tarefa proposta foi investigar o comportamento temporal sob outros aspectos, por exemplo, vendas de touros de categorias CL e LA. A formulação de novas tarefas ou a reformulação de tarefas já definidas é favorecida por dois motivos: primeiro, o usuário pode estabelecer correlações a partir da observação da visualização de *itemsets*, que indica possíveis correlações entre itens, dentro dos agrupamentos; segundo, a abordagem favorece a identificação de correlações pelo usuário a partir das regras obtidas durante o processo – não é necessário ir até o fim para obter todas as regras de associação e depois analisá-las.

Além disso, durante a exploração das regras, o especialista interrompeu o processo explorando as diversas relações entre regras ao analisar cada uma. A abordagem prevê a comparação visual de regras aos pares, recurso que permite que certas correlações entre grupos de itens envolvidos em duas regras possam ser claramente analisadas, como ocorreu durante a execução da **Tarefa 1**.

Este comportamento não-linear – mudar o valor do suporte mínimo durante o processo, reformular tarefas ou formular tarefas novas, interromper o processo para analisar duas ou mais regras – é inerente ao ser humano, e não é refletido em um processo convencional de obtenção de regras de associação, que é linear e fortemente direcionado, tanto na

etapa de obtenção de *itemsets* freqüentes quanto na etapa de geração de regras. A abordagem apresentada é um passo no sentido de adaptar o processo de extração de regras para oferecer mais suporte a este comportamento não-linear, característico de cenários exploratórios.

6.3 Estudos de caso em Cenário Exploratório

6.3.1 Primeiro Estudo – Cenário Exploratório

O objetivo deste estudo de caso foi avaliar o uso da abordagem interativa em um cenário exploratório, em que não se conhece a base de dados, com relação à determinação de agrupamentos de *itemsets* freqüentes e à determinação das tarefas a cumprir. Neste estudo, foi utilizada a base de dados do Orkut. Baseado na informação de participação de membros em comunidades, o usuário buscou encontrar comunidades correlacionadas.

Para este estudo, foram selecionados apenas os atributos MEMBRO_ID e nome da comunidade da base do Orkut, pois o objetivo era determinar associações entre comunidades, de acordo com os seus membros. Após a seleção inicial dos atributos, foi feita uma limpeza nos dados, com substituição de caracteres como: apóstrofes, aspas e & em HTML (*HyperText Markup Language*), emoticons e vírgulas. Foram também utilizados os nomes das comunidades para a sua identificação durante o processo. Nesta base, cada item corresponde a uma comunidade, e uma transação corresponde a um membro, de modo que cada transação (membro) contém uma lista de comunidades das quais participa.

Esta base era desconhecida do usuário (especialista em mineração), de modo que não foram determinadas previamente tarefas a serem realizadas. Decidiu-se primeiramente obter grupos de *itemsets* freqüentes, que indicariam grupos de comunidades mais interligadas. Depois, regras geradas a partir destes grupos seriam analisadas para identificar tipos de tarefas de análise potencialmente interessantes.

Inicialmente, foi estabelecido um suporte mínimo de 10%, depois de 5%, 2%, e por fim, valores absolutos de 20 ($\sim 1,25\%$) e de 10 ($\sim 0,62\%$), o que indica que foram consideradas na análise apenas as comunidades que têm 10 ou mais membros.

Dos 1.506 *itemsets* freqüentes de tamanho 2 obtidos após o passo 2 do algoritmo *Apriori*, foram excluídos 772, pois não despertaram interesse. Os *itemsets* excluídos contêm itens representando comunidades como “Eu ando com meus pés”, “Eu já fui um espermatozóide”, “Já desci escada pelo corrimão!”, “Quem não cola não sai da escola”, “Minha bagunça é organizada!”, “Saia do ORKUT e vá ler 1 LIVRO”, dentre outros. A partir dos *itemsets* freqüentes restantes, foram geradas 624 regras a partir de 14

agrupamentos (determinados pelo algoritmo *k*-médias – utilizando a métrica *CityBlock* – sobre a projeção obtida), das quais 390 foram consideradas interessantes, 100 óbvias e 108 irrelevantes (26 não foram avaliadas). O julgamento do interesse das regras foi feito pelo usuário, utilizando conhecimento prévio adquirido por cada um (a base de dados era desconhecida).

O mesmo foi feito para o próximo passo (o terceiro) do algoritmo *Apriori*. Foram obtidos 61 *itemsets* frequentes, a partir dos quais foram geradas 53 regras, sendo 34 interessantes, 5 óbvias e 9 irrelevantes (5 não foram avaliadas). Na **Figura 6.5(a)**, é mostrada a visualização dos agrupamentos de *itemsets* de tamanho 2, enquanto que a dos de tamanho 3 é mostrada na **Figura 6.5(b)**. Resumos das regras obtidas a partir dos agrupamentos de *itemsets* frequentes de tamanhos 2 e 3 são mostrados nas **Tabelas 6.10** e **6.11**, respectivamente.

Figura 6.5: Agrupamentos de *itemsets* frequentes obtidos a partir da base Orkut.

À medida que o conjunto de regras ia sendo explorado, determinados itens-alvo para as tarefas foram sendo estabelecidos, por exemplo:

- marca de produto (Ex.: Rip Curl, Billabong, Volcom etc.)
- programa de televisão (Ex.: Seu Madruga é Rei!!! (Chaves), The Simpsons - Brasil, THUNDERCATS etc.)
- clube/liga (Ex.: Futbol Club Barcelona, AC MILAN, Real Madrid CF etc.)
- personalidade (Ex.: Ronaldo, Ayrton Senna)
- comportamento (Ex.: Fico na NET até de Madrugada!, Eu Não Dirijo – Eu Piloto!)

Tabela 6.10: Número de regras obtidas de 2-*itemsets*, separadas por agrupamento.

Agrupamento	Total de regras	Regras interessantes	Regras óbvias	Regras irrelevantes	Regras não avaliadas
2-00	22	4	8	8	2
2-01	9	3	4	2	0
2-02	25	17	6	2	0
2-03	17	7	10	0	0
2-04	3	3	0	0	0
2-05	21	18	2	0	1
2-06	85	27	27	13	18
2-07	19	11	3	4	1
2-08	100	65	20	11	4
2-09	38	31	0	7	0
2-10	98	69	11	18	0
2-11	45	35	0	10	0
2-12	61	51	4	6	0
2-13	81	49	5	27	0
Total	624	390	100	108	26

Tabela 6.11: Número de regras obtidas de 3-*itemsets*, separadas por agrupamento.

Agrupamento	Total de regras	Regras interessantes	Regras óbvias	Regras irrelevantes	Regras não avaliadas
3-00	7	1	3	0	3
3-01	9	0	1	8	0
3-02	8	7	0	0	1
3-03	13	11	1	0	1
3-04	12	11	0	1	0
3-05	4	4	0	0	0
Total	53	34	5	9	5

- localidade (Ex.: São Paulo, Florianópolis)
- *hobby* (Ex.: Tuning, Viajando de carro)
- grupos musicais, cantores(as), estilos musicais (Ex.: Asa de águia, Chiclete com Banana, Tupac etc.)
- gastronomia (Ex.: Chocolate, Fondue)

A partir daí, foram selecionados grupos de regras contendo diferentes combinações destes itens-alvo. Dado o interesse do consumidor em uma determinada marca (ou grupo, escritor etc.), busca-se por exemplo, identificar outras preferências (marcas, grupos, escritores etc.) relacionadas. Estes grupos de regras são mostrados a seguir, sendo que são mostradas nas **Tabelas 6.12 a 6.25** apenas as regras melhor ranqueadas em cada caso.

Uma das tarefas exploratórias consistiu em comparar grupos de regras contendo um item em foco, visando descobrir quais os itens mais relacionados a ele. Este tipo de tarefa pode ser realizado com o módulo RulEx.

Tabela 6.12: Regras envolvendo duas marcas de produtos.

Cluster	Antecedente	Consequente	Suporte	Confiança	<i>Lift</i>
2-08	Cavalera	Puma	18	36,73%	5,514
2-08	Osklen	Puma	10	38,46%	5,773
2-10	Cavalera	Adidas	22	44,90%	4,622
2-10	Billabong	Adidas	19	38,78%	3,992
2-11	Volcom	Nike	17	45,95%	7,607
2-11	Quiksilver	Nike	20	33,33%	5,519
2-12	Porsche	Dodge_charger	10	35,71%	7,751
2-10	Coca_Cola	Adidas	16	35,56%	3,660
2-12	Fanta_Uva	Dodge_charger	12	36,36%	7,892

Tabela 6.13: Regras envolvendo três marcas de produtos.

Cluster	Antecedente	Consequente	Suporte	Confiança	<i>Lift</i>
3-03	Puma,Quiksilver	Nike	11	61,11%	10,118
3-03	Adidas,Cavalera	Puma	13	59,09%	8,869
3-03	Adidas,FERRARI	Puma	10	62,50%	9,381
3-05	Quiksilver,Nike	Volcom	11	55,00%	23,873

Tabela 6.14: Regras envolvendo um clube/liga e uma marca.

Cluster	Antecedente	Consequente	Suporte	Confiança	<i>Lift</i>
2-08	Futbol_Club_Barcelona	Puma	10	33,33%	5,003
2-10	Futbol_Club_Barcelona	Adidas	15	50,00%	5,147
2-10	AC_MILAN	Adidas	13	54,17%	5,576
2-10	Real_Madrid_CF	Adidas	12	54,55%	5,615
2-10	UEFA_Champions_League	Adidas	22	59,46%	6,121
2-10	NBA_Basketball	Adidas	11	50,00%	5,147

Tabela 6.15: Regras envolvendo uma marca e uma personalidade.

Cluster	Antecedente	Consequente	Suporte	Confiança	<i>Lift</i>
2-10	Alexandre_Pato	Adidas	11	40,74%	4,194
2-10	Ronaldo	Adidas	10	37,04%	3,813
2-13	FERRARI	Ayrton_Senna	35	49,30%	3,862
2-13	Playboy	Ayrton_Senna	20	45,45%	3,561
2-13	Meu_celular_é_Vivo	Ayrton_Senna	13	46,43%	3,637
2-13	Johnnie_Walker	Ayrton_Senna	21	35,00%	2,742
2-13	Cerveja_Bohemia	Ayrton_Senna	19	33,93%	2,658
2-13	Porsche	Ayrton_Senna	10	35,71%	2,798
2-13	Skol Beats	Ayrton_Senna	10	45,45%	3,561

Tabela 6.16: Regras envolvendo uma marca e um *hobby*.

Cluster	Antecedente	Consequente	Suporte	Confiança	<i>Lift</i>
2-03	BMW_Enthusiasts	Viajando_de_Carro	12	37,50%	6,922
2-06	Lamborghini	Tuning	16	34,78%	4,297
2-06	iPod	Tuning	10	38,46%	4,751
2-06	BMW_Enthusiasts	Tuning	11	34,38%	4,247
2-08	Coca_Cola	Churrasco	16	35,56%	4,965

Outro tipo de tarefa consistiu em analisar o relacionamento indireto (se A e B, e B e C são associados, então A e C também são associados?) entre itens contidos em um grupo de

Tabela 6.17: Regras envolvendo uma marca e um cantor/grupo/estilo musical.

Cluster	Antecedente	Conseqüente	Suporte	Confiança	Lift
2-08	Marcelo_D2	Puma	15	35,71%	5,360
2-10	Eminem_	Adidas	19	42,22%	4,347
2-10	Racionais_Mcs	Adidas	11	33,33%	3,432
2-10	Linkin_Park_-_Brasil	Adidas	10	30,30%	3,120
2-10	Eminem_	Adidas	19	42,22%	4,347
2-10	Marcelo_D2	Adidas	18	42,86%	4,412
2-10	Snoop_Dogg	Adidas	14	45,16%	4,649
2-10	Hip_Hop	Adidas	15	34,09%	3,510
2-10	Reggae	Adidas	11	36,67%	3,775
2-11	Marcelo_D2	Nike	14	33,33%	5,519

Tabela 6.18: Regras envolvendo dois programas de TV.

Cluster	Antecedente	Conseqüente	Suporte	Confiança	Lift
2-02	Os_Trapalhães	Eu_amo_o_Seu_Madruga_-_CHAVES	15	41,67%	7,119
2-05	Seu_Madruga_é_Rei!!!	The_Simpsons_-_Brasil	12	42,86%	6,748
2-08	Carrossel	Caverna_do_Dragão	10	37,04%	5,508

Tabela 6.19: Regras envolvendo um programa de TV e uma personalidade.

Cluster	Antecedente	Conseqüente	Suporte	Confiança	Lift
2-05	Ronaldo	The_Simpsons_-_Brasil	10	37,04%	5,832
2-13	Smallville_Brasil	Ayrton_Senna	10	34,48%	2,701
2-13	Zacarias	Ayrton_Senna	10	37,04%	2,902
2-13	Carrossel	Ayrton_Senna	10	37,04%	2,902
2-13	Chaves	Ayrton_Senna	26	34,67%	2,716
2-13	Os_Trapalhães	Ayrton_Senna	14	38,89%	3,047
2-13	Doug_Funnie	Ayrton_Senna	14	33,33%	2,611

Tabela 6.20: Regras envolvendo um programa de TV e um cantor/grupo/estilo musical.

Cluster	Antecedente	Conseqüente	Suporte	Confiança	Lift
2-02	Angra	Eu_amo_o_Seu_Madruga_-_CHAVES	10	40,00%	6,834
2-09	Os_Trapalhães	Mamonas_Assassinas	13	36,11%	4,640
2-09	RockGoLMTV	Mamonas_Assassinas	10	35,71%	4,589
2-12	Harry_Potter_Brasil	Jack_Johnson	10	33,33%	4,911

Tabela 6.21: Regras envolvendo um comportamento e um *hobby*.

Cluster	Antecedente	Conseqüente	Suporte	Confiança	Lift
2-06	Fico_na_NET_até_de_Madrugada!	Tuning	26	35,62%	4,400
2-06	Eu_Não_Dirijo_-_Eu_Piloto!	Tuning	12	38,71%	4,782
2-06	NEED_FOR_SPEED_WORLD	Tuning	10	45,45%	5,615
2-13	Eu_Amo_Carros_Rebaixados	EU_AMO_VELOCIDADE	12	80,00%	19,467

regras, buscando comparar a intensidade das associações entre estes itens. Por exemplo, as associações de Adidas com Puma e com Nike são mais fortes que a associação entre Puma e Nike em si, como indicam as regras na **Tabela 6.26**. Outro exemplo de análise de

Tabela 6.22: Regras envolvendo uma localidade e outro item.

Cluster	Antecedente	Conseqüente	Suporte	Confiança	<i>Lift</i>
2-10	Baladas_Vila_Olímpia	São_Paulo	13	44,83%	5,624
2-10	ZECA_PAGODINHO	São_Paulo	10	34,48%	4,327
2-10	Balneário_Camboriu	Adidas	11	35,48%	3,653
2-13	Florianopolis	Ayrton_Senna	11	39,29%	3,078

Tabela 6.23: Regras envolvendo um *hobby* e um cantor/grupo/estilo musical.

Cluster	Antecedente	Conseqüente	Suporte	Confiança	<i>Lift</i>
2-06	Eminem_	Tuning	19	42,22%	5,216
2-06	Hip_Hop	Tuning	16	36,36%	4,492
2-06	Snoop_Dogg	Tuning	11	35,48%	4,384

Tabela 6.24: Regras envolvendo dois grupos musicais.

Cluster	Antecedente	Conseqüente	Suporte	Confiança	<i>Lift</i>
2-02	Angra	Iron_Maiden_Brasil	11	44,00%	18,596
2-10	Blink_182	Green_Day	10	52,63%	31,306
2-10	The_Offspring	Green_Day	10	43,48%	25,862
2-12	Bob_Marley	Jack_Johnson	20	33,90%	4,995

Tabela 6.25: Regras envolvendo duas marcas e outro item.

Cluster	Antecedente	Conseqüente	Suporte	Confiança	<i>Lift</i>
3-02	Adidas,?_Counter-Strike_Brasil_(CS)	Nike	10	55,56%	9,198
3-02	Adidas,Tuning	Nike	14	46,67%	7,726
3-02	Puma,Tuning	Nike	10	52,63%	8,714
3-02	Eminem_,Adidas	Nike	10	52,63%	8,714
3-02	Adidas,Fico_na_NET_até_de_Madrugada!	Nike	10	47,62%	7,884
3-03	Puma,Ayrton_Senna	Adidas	14	70,00%	7,206
3-03	The_Simpsons_-_Brasil,Nike	Puma	10	76,92%	11,546
3-03	Puma,UEFA_Champions_League	Adidas	10	90,91%	9,359
3-03	Ayrton_Senna,Nike	Adidas	14	63,64%	6,551
3-03	Adidas,Marcelo_D2	Puma	11	61,11%	9,172
3-03	Puma,Tuning	Adidas	12	63,16%	6,502
3-03	Adidas,Durmo_com_meu_celular_do_lado	Puma	13	52,00%	7,805
3-03	Happy_Tree_Friends,Puma	Adidas	10	66,67%	6,863
3-04	Puma,Ayrton_Senna	FERRARI	11	55,00%	12,441
3-04	Volkswagen_do_Brasil,FERRARI	Ayrton_Senna	10	76,92%	6,026
3-04	Adidas,FERRARI	Ayrton_Senna	11	68,75%	5,386

relacionamentos indiretos é mostrado na **Figura 6.6**. Billabong e Quiksilver (dois itens) se encontram no centro dos relacionamentos. Conectados a eles estão Rip_Curl (apenas a Billabong e a Reef), e Reef e Volcom (conectados a ambos). Periféricamente, estão OAKLEY e Hang_Loose (ligados a Billabong) e Element (ligado a Volcom). As regras envolvidas são mostradas na **Tabela 6.27**. Na representação esquemática, cada elipse corresponde a um item, descrito pelo rótulo, e cada aresta corresponde a uma ligação, cuja intensidade é indicada pela sua espessura (quanto mais espessa é a aresta, maior é a intensidade), determinada pelos valores de suporte, confiança e *lift* da regra.

Tabela 6.26: Conjunto de regras envolvendo três marcas específicas de produtos.

Cluster	Antecedente	Conseqüente	Suporte	Confiança	Lift
2-08	Nike	Puma	41	42,27%	6,344
2-10	Puma	Adidas	50	46,73%	4,811
2-10	Nike	Adidas	44	45,36%	4,670

Tabela 6.27: Conjunto de regras envolvendo oito marcas específicas de produtos.

Cluster	Antecedente	Conseqüente	Suporte	Confiança	Lift
2-06	Rip_Curl	Billabong	17	73,91%	24,225
2-06	Volcom	Quiksilver	22	59,46%	15,915
2-06	Element	Volcom	14	82,35%	35,746
2-06	Billabong	Quiksilver	25	51,02%	13,656
2-06	Reef	Billabong	15	53,57%	17,558
2-06	Volcom	Billabong	17	45,95%	15,059
2-06	Hang_Loose	Billabong	10	100,00%	32,776
2-06	Rip_Curl	Reef	11	47,83%	27,432
2-06	Reef	Quiksilver	14	50,00%	13,383
2-06	OAKLEY	Billabong	10	22,22%	7,283

Figura 6.6: Representação esquemática das associações entre os itens.

6.3.2 Segundo Estudo – Tarefa Específica

O objetivo deste estudo foi comparar a realização de tarefas de obtenção de regras de associação, com a abordagem, em um cenário em que o minerador já fez uma pré-exploração, com um cenário em que o minerador ainda não explorou a base de dados.

Como ponto de partida, inicialmente foi designada uma tarefa específica a ser cumprida, de encontrar regras específicas envolvendo marcas de carro e de roupas. Então, cada um dos dois usuários (ambos especialistas em mineração) foi designado a realizá-la com o auxílio do sistema, que suporta a abordagem.

Inicialmente, o primeiro usuário ajustou o suporte mínimo para o valor 35 ($\sim 2,18\%$), mudando posteriormente para 10 ($\sim 0,62\%$). Visto que a tarefa é específica, o minerador escondeu alguns *itemsets* que julgou não corresponderem aos itens-alvo.

No passo 2, o usuário selecionou *itemsets* com um item envolvendo carros e com um item envolvendo marcas de roupas. Após cada seleção, o usuário gerou regras envolvendo os *itemsets* selecionados. Foram geradas 310 regras de tamanho 2, das quais apenas 7, mostradas na **Tabela 6.28** foram selecionadas.

Tabela 6.28: Conjunto de regras envolvendo diversas marcas de produtos.

Antecedente	Consequente	Suporte	Confiança	Lift
Peugeot_206_Brasil	Puma	10	9,80%	1,472
Peugeot_206_Brasil	Adidas	11	10,78%	1,110
FERRARI	Puma	19	26,76%	4,017
Lamborghini	Adidas	14	30,43%	3,133
FERRARI	Adidas	16	22,54%	2,320
Dodge_charger	Adidas	12	16,21%	1,669
Volkswagen_do_Brasil	Adidas	16	13,01%	1,339

O mesmo foi feito no passo 3 (para os *itemsets* de tamanho 3), sendo geradas 41 regras, das quais apenas uma foi selecionada: *Adidas, Nike* $\rightarrow$ *Tuning* (suporte 14, confiança 31,82%, lift 3,931).

O segundo minerador, por outro lado, utilizou inicialmente um suporte alto – 169 ($\sim 10\%$), não aparecendo nenhum *itemset*. O procedimento foi repetido para suporte 85 ($\sim 5\%$) e 32 ($\sim 2\%$).

Iniciando a exploração dos *itemsets* pelos agrupamentos maiores, o usuário identificou *itemsets* envolvendo carro com carro, roupa com roupa e carro com música. Pela visualização, ele correlacionou indiretamente *Adidas* a *Ferrari* e a *Volkswagen*, por meio do item *Ayrton Senna*.

Não satisfeito com o suporte mínimo indicado, o minerador mudou novamente este valor para 16 ($\sim 1\%$). Desta vez, ele eliminou *itemsets* contendo alguns itens não buscados. Explorando pela lista de itens, o usuário selecionou *Adidas* e *Nike*, encontrando apenas a regra (1) da **Tabela 6.29**. Continuando a exploração, o minerador encontrou as regras (2) e (3) da mesma tabela.

Tabela 6.29: Conjunto de regras envolvendo diversas marcas de produtos.

ID	Antecedente	Consequente	Suporte	Confiança	Lift
(1)	FERRARI	Nike	17	23,94%	3,964
(2)	Puma	Tuning	19	17,76%	2,194
(3)	FERRARI	Puma	19	26,76%	4,017

O usuário comparou então as duas regras envolvendo o item *FERRARI* e inferiu que, nessa base, Puma está mais relacionado à esportividade, a um grupo mais jovem de comportamento do que *Nike*.

O minerador ainda analisou os *itemsets* contendo os itens *Volkswagen_Gol*, *Volkswagen_Voyage* e *Volkswagen* (agrupados na lista), mas não achou regras ligadas a roupas. Fez o mesmo para outros itens representando outras marcas de carro (*Peugeot*, *Porsche*, *Lamborghini*), mas não achou regras envolvendo roupas. Por fim, analisou os *itemsets* contendo itens representando outras marcas de roupa (*Volcom*, *Quiksilver*, *Billabong*, *Rip Curl*), mas não achou regras envolvendo carros.

No início do processo, o primeiro usuário adotou apenas dois valores para o suporte mínimo, indicando inicialmente 35 ($\sim 2,18\%$) e mudando posteriormente para 10 ($\sim 0,62\%$). O segundo, por outro lado, utilizou quatro valores distintos na determinação do valor do suporte – inicialmente, indicou 169 ($\sim 10\%$), mudando para 85 ($\sim 5\%$), 32 ($\sim 2\%$), e posteriormente 16 ($\sim 1\%$), valor que julgou suficiente. Para fazer este julgamento, os mineradores utilizaram a visualização de *itemsets* e a lista de itens, de modo a obter uma quantidade razoável deles para exploração. Como esperado, o primeiro minerador, que já havia explorado anteriormente, ajustou um menor número de vezes o valor do suporte mínimo, por já ter uma visão geral do conjunto de *itemsets* que poderia ser obtido.

Em ambos os casos, o minerador focou em itens específicos, por meio da lista de itens mostrada na interface do sistema. A lista de itens diminui a carga cognitiva à medida que auxilia o minerador na seleção de itens-alvo. Focando nos itens, os mineradores tiveram atitudes distintas. O primeiro foi mais direto, selecionando os itens, individualmente, e gerando as regras correspondentes a eles, enquanto que o segundo buscou estabelecer correlações entre os itens por meio do seu posicionamento na visualização. Não tendo uma visão geral dos *itemsets*, o segundo minerador utilizou os recursos disponíveis para uma exploração mais detalhada, ao contrário do primeiro.

Em ambos os casos, a tarefa pôde ser completada, e em pouco tempo. Este é um indício de que a abordagem pode ser usada nestes dois cenários – em que o minerador já possui uma visão geral, e em que o minerador desconhece totalmente a base de dados – sendo que no segundo a abordagem é mais intensamente utilizada.

No primeiro estudo, foram observadas características presentes nos estudos anteriores, como o auxílio na determinação do limiar de suporte mínimo, que precisou de algumas tentativas (cerca de 5 ou 6), evitando que diversas execuções completas do processo fossem realizadas. Além disso, foi obtido um conjunto compacto com 677 regras (624 de tamanho 2 e 53 de tamanho 3), das quais 424 (390 de tamanho 2 e 34 de tamanho 3) foram consideradas interessantes, 105 (100 de tamanho 2 e 5 de tamanho 3) óbvias, e 117 (108

de tamanho 2 e 9 de tamanho 3) óbvias (31 – 26 de tamanho 2 e 5 de tamanho 3 – não foram avaliadas).

O primeiro estudo também indica que é possível definir e realizar tarefas em um cenário exploratório a partir das regras geradas dentro de agrupamentos. Partindo de um cenário inicialmente desconhecido do minerador, é possível direcionar o processo de descoberta em sub-tarefas, a partir da segmentação dos *itemsets* freqüentes, que dão indícios de itens possivelmente associados. Estas sub-tarefas guiam o minerador na determinação das tarefas a serem executadas.

No segundo estudo, em que os mineradores deveriam extrair regras com objetivo bem específico, a tarefa foi cumprida e em pouco tempo, indicando que a abordagem pode efetivamente auxiliar o usuário. Primeiramente, a lista de itens ajudou na determinação de itens-alvo, diminuindo a carga cognitiva exigida para o cumprimento da tarefa. Além disso, o posicionamento dos *itemsets* na visualização auxilia na exploração dos *itemsets*, favorecendo a inferência de correlações diretas e indiretas.

6.4 Considerações Finais

Neste capítulo, foram descritos diversos estudos buscando mostrar os recursos em uso por usuários reais, tanto para situações em que há um especialista quanto para cenários exploratórios. Em ambos os casos, a mineração foi realizada satisfatoriamente, com algumas vantagens em relação a uma abordagem convencional.

Primeiramente, nos estudos com a base PMGRN, o especialista do domínio destacou, dentre as características da abordagem, a importância de comparação de valores de medidas de regras. Segundo o especialista, a comparação é um instrumento útil na escolha de regras como sendo interessantes, pois permite comparar regras com outras relacionadas a ela. Além disso, no primeiro estudo com a base Orkut, o minerador destacou a mudança dinâmica do suporte mínimo, fazendo uso intenso desta característica da abordagem para determinar o seu valor de maneira adequada, observando as visualizações que foram sendo geradas. No segundo estudo com a base Orkut, o segundo minerador utilizou a visualização para inferir correlações entre itens, de acordo com o seu posicionamento na representação visual, o que indica a sua importância no processo. Estas inferências foram posteriormente validadas ou descartadas, o que enriqueceu a exploração da base de dados.

No **Capítulo 7**, serão apresentadas as conclusões finais e as contribuições deste trabalho de doutoramento. Também são discutidas as limitações da abordagem, e são apresentados direcionamentos para trabalhos futuros e para melhoramento do Sistema I₂E.

Conclusões

A tarefa de obtenção de regras de associação apresenta diversas dificuldades para os mineradores, relacionadas à configuração de parâmetros de entrada, ao grande número de regras geradas, e à sua compreensão. Diversas abordagens tratam estes problemas utilizando meios de flexibilizar o processo, e de explorar os conjuntos de regras, com ou sem apoio de representações visuais.

Nesse contexto, este trabalho apresentou uma abordagem apoiada na extração interativa de *itemsets* frequentes e na exploração direcionada de regras. A primeira provê mecanismos para execução interativa do *Apriori*, exploração visual de *itemsets* frequentes e extração de regras baseada nesses *itemsets*. A segunda etapa permite a exploração de regras de associação por pares e comparação de suas medidas de interesse. Esta abordagem foi integrada a um sistema computacional de apoio ao processo, chamado I₂E.

Foram realizados diversos estudos para avaliar a abordagem, apoiados pelo uso do sistema. Estes estudos indicaram que ela é adequada para a obtenção de regras de associação, pois ajuda a obter conjuntos de regras compactos, com boa cobertura de *itemsets*, além de possibilitar o desenvolvimento de um processo de maneira não-linear que é inerente ao ser humano, tanto para cenários exploratórios quanto para cenários em que há previamente um direcionamento para o processo.

A seguir, são apresentadas as contribuições deste trabalho apresentado, as limitações da abordagem e direcionamentos para trabalhos futuros.

7.1 Contribuições

O desenvolvimento deste trabalho foi guiado por uma hipótese geral, a de que é possível inserir o minerador/especialista no contexto de extração de regras de associação, com o objetivo de facilitar a realização de tarefas desta natureza, por acompanhar o processo de perto, e de permitir que resultados mais ricos sejam obtidos, e por inferir correlações durante a exploração. O minerador assume um papel ativo no monitoramento e direcionamento do processo, alterando parâmetros de entrada e do algoritmo analítico durante a sua execução. Para isso, a abordagem apresentada utiliza recursos para a visualização e seleção de *itemsets* freqüentes e de valores de medidas de regras de associação. Tomando como ponto de partida esta hipótese geral, duas hipóteses específicas foram apresentadas no **Capítulo 1** para guiar o desenvolvimento deste trabalho.

A primeira diz que é possível facilitar (para o usuário) o processo de obtenção de regras de associação, auxiliando-o a determinar os parâmetros do algoritmo e extraíndo conjuntos mais reduzidos de regras, como ponto de partida para a exploração do conhecimento. Os estudos (com especialista, apresentados na seção **Seção 6.2** e em cenários exploratórios, apresentados na seção **Seção 6.3**) deram indícios de que a determinação do nível de suporte mínimo foi feita com menos esforço computacional do que de uma maneira convencional, *ad hoc*. Nestas avaliações, foram utilizadas cerca de três a quatro configurações diferentes para o suporte mínimo. Nestes estudos, somente uma execução integral do *Apriori* foi necessária, sendo preciso repetir apenas os dois primeiros passos (obtenção dos itens freqüentes e a obtenção dos *itemsets* freqüentes de tamanho 2). A execução de etapas do *Apriori* propicia ao usuário um desenvolvimento mais cadenciado, de modo que ele pode monitorar as ações sendo executadas durante determinados *check-points* (momentos entre as etapas). Em uma abordagem convencional, seriam necessárias diversas execuções completas do *Apriori* (cerca de quatro a seis), cujos resultados seriam enormes conjuntos de regras, difíceis de manipular. Por enorme, entende-se por grandes o suficiente para tornar a atividade de exploração desestimulante. Além disso, nas avaliações da **Cobertura de Itemsets** e da **Política de Filtragem**, verificou-se que, com a abordagem interativa, foram obtidos conjuntos de regras com maior cobertura de *itemsets*, e mais compactos, respectivamente, em comparação com outras abordagens existentes, preservando a diversidade de regras e a cobertura de *itemsets*.

A segunda hipótese diz que é possível indicar ao usuário caminhos de exploração de regras em um cenário exploratório e aumentar a sua capacidade de interpretação dos resultados, com o apoio de representações visuais adequadas. Uma vez engajado com a realização da tarefa, o usuário pode formular novos caminhos para descoberta de

conhecimento por meio da indução de novas correlações entre itens. O segundo estudo de caso **Com Especialista do Domínio** e os estudos **Em Cenários Exploratórios** dão indícios de que o usuário pode induzir correlações entre itens, baseado na exploração visual de *itemsets* freqüentes. Durante a realização da primeira tarefa da segunda avaliação com especialista, o usuário percebeu que, ao invés de utilizar dois itens, poderia utilizar três, e gerar regras mais específicas do que era esperado inicialmente. Desta maneira, o usuário interrompeu a tarefa e prosseguiu, obtendo regras contendo três itens. Na segunda parte do estudo de caso em cenário exploratório, um dos mineradores utilizou a representação visual de *itemsets* para inferir correlações entre itens não explicitamente correlacionados, para posteriormente validá-las ou descartá-las. Estas análises podem ser feitas diretamente na exploração de *itemsets* posicionados de maneira aproximada (os itens que não são em comum entre eles podem ser combinados em outros *itemsets*, provocando o surgimento de inferências e indicando novas direções na exploração) ou por meio dos agrupamentos automáticos de *itemsets* – e por conseguinte, de regras, cuja seleção indica quais os itens que mais ocorrem (os itens podem ser recombinaados em outros *itemsets*, provocando o surgimento de inferências e indicando novas direções na exploração). Além disso, nos estudos de caso com usuário, a comparação de regras auxiliou na determinação do interesse das regras, aumentando o poder de análise dos usuários, enriquecendo a análise.

Desta maneira, a metodologia proposta contribui para:

- Aumentar a clareza da execução do algoritmo *Apriori*, permitindo que os passos sejam avançados ou retrocedidos, de modo a fazer com que eles sejam monitorados mais facilmente pelo usuário.
- Fornecer uma visualização significativa de conjuntos de *itemsets* freqüentes, de modo que a sua localização na representação visual mapeie relações de similaridade/dissimilaridade entre eles, permitindo que o usuário possa inferir correlações entre itens contidos nestes *itemsets* e tomando novas direções na exploração.
- Fornecer uma extração de conjuntos compactos e significativos de regras de associação, baseada em *itemsets* e orientada por agrupamentos.
- Fornecer uma exploração de regras por comparação em pares, de modo a facilitar a análise que envolve regras de associação relacionadas por algum tipo de critério.
- Facilitar o processo de comparação de regras por pares, por meio de uma visualização de regras usando coordenadas paralelas que destaca as diferenças de valores entre duas regras e entre uma regra e um conjunto de regras.

A lista de artigos publicados e em preparação é mostrada no **Apêndice B**. O sistema e o seu manual podem ser encontrados em `infoserver.lcad.icmc.usp.br/infovis2/I2E`.

7.2 Limitações

Apesar de representar um avanço no sentido de prover um papel ativo ao usuário na realização de tarefas de obtenção de regras de associação, com apoio de técnicas de visualização de informação, a abordagem proposta apresenta algumas limitações do ponto de vista global:

- Ela provê meios para que o minerador identifique regras potencialmente interessantes com uma quantidade menor de tentativas (provendo visualização dos *itemsets*) e para que o processo como um todo seja menos cansativo, se comparado a uma abordagem convencional, porém não resolve por completo o problema de definição do suporte mínimo “ideal”, apenas auxilia o minerador a ajustá-lo.
- Ela auxilia na obtenção de regras envolvendo itens com suporte relativamente baixo, porém não trata explicitamente o problema de obtenção de *itemsets* raros, isto é, *itemsets* que não ocorrem com frequência.
- Ela auxilia na exploração de conjuntos de regras, visando a escolha de regras interessantes, porém não aborda explicitamente o problema da interpretação de regras.
- Ela aborda a obtenção de *itemsets* frequentes, não podendo ser aplicada à obtenção de padrões sequenciais, isto é, de associações que ocorrem em sequência no tempo.

A construção do sistema de apoio à abordagem foi guiada por várias decisões de implementação que introduzem algumas limitações específicas de ordem prática:

- A visualização de *itemsets* auxilia o minerador na determinação de agrupamentos de *itemsets* (dividindo o espaço de busca e favorecendo o estabelecimento de correlações entre itens). Entretanto, como as projeções são atualizadas a cada passo, ela não garante que *itemsets* envolvendo os mesmos itens, em passos diferentes do processo, se mantenham na mesma região, o que requer que o usuário crie um mapa mental diferente para cada passo.
- O incremento do valor do suporte mínimo é viável e não requer uma re-execução do *Apriori*, já que as informações sobre todos os *itemsets* são mantidas na estrutura de dados. Por outro lado, se o valor for reduzido, o processo precisa ser reiniciado,

a não ser que uma estrutura de dados armazene *a priori* informações relativas a todos os *itemsets* que podem ser gerados, a partir do menor limiar de suporte possível. Esta política, entretanto, seria onerosa em termos de consumo de memória e processamento.

- A extração de vetores de características que descrevem os *itemsets*, e que são a base para a visualização de *itemsets*, considerou que todas as relações de pertinência de itens, em seus respectivos *itemsets*, têm o mesmo peso. Se considerarmos que o nível de interesse do usuário em relação aos itens é variável, o peso poderia variar, conforme este interesse, e isso poderia se refletir na visualização de *itemsets*.
- A projeção de conjuntos de *itemsets* freqüentes de tamanho pequeno nem sempre proporciona uma distinção clara de grupos de *itemsets*. Isto se deve ao próprio fato de os *itemsets* freqüentes serem pequenos, não fornecendo muitas informações nos vetores de características, que os aproximem/distanciem uns dos outros.
- Na geração de regras, é utilizada a posição no *ranking* como critério para selecionar a regra representativa para cada *itemset*. Entretanto, em algumas ocasiões, há empate, e a escolha se dá pela ordem de geração. Uma alternativa é dar ao minerador a opção de escolher as regras pelo formato (antecedente máximo e conseqüente mínimo ou o oposto) quando há empate.
- Na exploração de regras por pares, inicia-se a exploração por um conjunto de regras. Entretanto, pode ser do interesse do usuário comparar as regras das tabelas de regras relacionadas com regras relacionadas a estas. O Sistema I₂E não prevê este caso.

7.3 Trabalhos Futuros

A seguir, são listadas algumas propostas para desenvolvimento de trabalhos futuros, visando tratar os problemas levantados nesta tese.

- Elaboração de uma metodologia para exploração visual de *itemsets* com suporte menor que o suporte mínimo determinado. Pode-se armazenar informações sobre *itemsets* que possuem suporte um pouco abaixo do suporte mínimo, e mostrar as correlações deles com os demais (freqüentes). Os *itemsets* obtidos a partir daí poderiam ser formados por todos os *itemsets* (freqüentes ou não-freqüentes).
- Elaboração de uma abordagem para ajuste do suporte até convergir (permitindo aumentar e diminuir até limites definidos em tentativas anteriores).

- Investigação da possibilidade de o usuário definir múltiplos níveis de suporte para diferentes grupos de *itemsets*.
- Aplicação de representações visuais ao problema de obtenção de *itemsets* raros. Pode-se, por exemplo, identificar e visualizar os *itemsets* que não ocorrem com frequência nas bases de dados, destacando-os dos demais.
- Aplicação de representações visuais ao problema de obtenção de padrões sequenciais, destacando visualmente as relações de causa e consequência entre os itens das bases de dados.
- Extensão do modelo, com o objetivo de permitir o trabalho colaborativo entre especialistas de domínio. Neste modelo, dados de exploração são compartilhados, permitindo que cada especialista possa redirecionar o processo de descoberta, baseado nas ações do outro, dividindo as tarefas, enriquecendo a exploração.

Além de tratar as limitações dos módulos do sistema, são listadas a seguir outras propostas que podem contribuir para melhorar a abordagem interativa.

- Elaboração de uma metáfora visual baseada em um relevo, no qual a altura de montanhas representa o tamanho dos *itemsets*. Nesta representação, poderia-se construir um terreno de maneira incremental à medida que o *Apriori* é executado, de modo a representar o avanço no processo, ao mesmo tempo que preserva o posicionamento dos *itemsets*.
- Elaboração de outros critérios para definição do vetor de características utilizado para o cálculo da distância entre os *itemsets*, por exemplo, levando em consideração as frequências de outros *itemsets* que também contêm os itens que eles contêm.
- Concepção de uma estrutura de indexação como suporte para a realização de cada tipo de consulta (relação de comparação) para obtenção de regras relacionadas.
- Estudo dos principais tipos de análise feitas pelo especialista, visando a elaboração de novas relações de comparação entre regras de associação.
- Combinação da política de filtragem de regras com outras já existentes, por exemplo, utilizando a eliminação de regras redundantes (discutidas no levantamento bibliográfico) e o agrupamento de regras, de acordo com o seu conteúdo (antecedente e consequente) ou utilizando *itemsets* maximais.

-
- Incorporação ao sistema de visualização como apoio para determinação do limiar de corte.
 - Investigação de heurísticas para determinar automaticamente um limiar/percentual adequado para o limiar de corte.

Referências Bibliográficas

- AGGARWAL, C. C. Towards effective and interpretable data mining by visual interaction. *SIGKDD Explorations*, v. 3, n. 2, p. 11–22, 2002.
- AGGARWAL, C. C.; YU, P. S. Mining large itemsets for association rules. *IEEE Data Engineering Bulletin*, v. 21, n. 1, p. 23–31, 1998a.
- AGGARWAL, C. C.; YU, P. S. Online generation of association rules. In: *Proceedings of International Conference on Data Engineering*, Orlando, FL, USA, 1998b, p. 402–411.
- AGGARWAL, C. C.; YU, P. S. A new approach to online generation of association rules. *IEEE Transactions on Knowledge and Data Engineering*, v. 13, n. 4, p. 527–540, 2001.
- AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. N. Mining association rules between sets of items in large databases. In: *Proceedings of ACM International Conference on Management of Data*, Washington, DC, USA, 1993, p. 207–216.
- ANKERST, M. *Visual data mining*. Tese de Doutoramento, Fakultät für Mathematik und Informatik der Ludwig-Maximilians-Universität München, 2000.
- ANKERST, M.; KEIM, D. A.; KRIEGEL, H.-P. ‘Circle Segments’: A technique for visually exploring large multidimensional data sets. In: *Proceedings of IEEE Visualization*, San Francisco, CA, USA, 1996.
- ASHRAFI, M. Z.; TANIAR, D.; SMITH, K. A. A new approach of eliminating redundant association rules. In: *Proceedings of International Conference on Database and Expert Systems Applications*, Zaragoza, Spain, 2004, p. 465–474.
- ASIMOV, D. The grand tour: a tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing archive*, v. 6, n. 1, p. 128–143, 1985.

- ASUNCION, A.; NEWMAN, D. J. UCI Machine Learning Repository. 2007.
Disponível em archive.ics.uci.edu/ml
- BASTIDE, Y.; PASQUIER, N.; TAOUIL, R.; STUMME, G.; LAKHAL, L. Mining minimal non-redundant association rules using frequent closed itemsets. In: *Proceedings of International Conference on Computational Logic*, London, UK: Springer-Verlag, 2000, p. 972–986.
- BEDDOW, J. Shape coding of multidimensional data on a microcomputer display. In: *Proceedings of IEEE Visualization*, San Francisco, CA, USA, 1990, p. 238–246.
- BESHERS, C.; FEINER, S. AutoVisual: Rule-based design of interactive multivariate visualizations. *IEEE Computer Graphics and Applications*, v. 13, n. 4, p. 41–49, 1993.
- BLANCHARD, J.; GUILLET, F.; BRIAND, H. Exploratory visualization for association rule rummaging. In: *Proceedings of International Workshop on Multimedia Data Mining*, Washington, DC, USA, 2003a, p. 107–114.
- BLANCHARD, J.; GUILLET, F.; BRIAND, H. A user-driven and quality-oriented visualization for mining association rules. In: *Proceedings of IEEE International Conference on Data Mining*, Melbourne, FL, USA, 2003b, p. 493–496.
- BLANCHARD, J.; GUILLET, F.; BRIAND, H. Interactive visual exploration of association rules with rule-focusing methodology. *Knowledge and Information Systems*, v. 13, n. 1, p. 43–75, 2007a.
- BLANCHARD, J.; PINAUD, B.; KUNTZ, P.; GUILLET, F. A 2D-3D visualization support for human-centered rule mining. *Computers and Graphics*, v. 31, n. 3, p. 350–360, 2007b.
- BRUZZESE, D.; BUONO, P. Combining visual techniques for association rules exploration. In: *Proceedings of Conference on Advanced Visual Interfaces*, Gallipoli, Italy, 2004, p. 381–384.
- CARVALHO, V. O. *Generalização no processo de mineração de regras de associação*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, 2007.
- CARVALHO, V. O.; REZENDE, S. O.; CASTRO, M. Obtaining and evaluating generalized association rules. In: *Proceedings of International Conference on Enterprise Information Systems*, Funchal, Madeira, Portugal, 2007, p. 310–315.

- CHAKRAVARTHY, S.; ZHANG, H. Visualization of association rules over relational DBMSs. In: *Proceedings of ACM Symposium on Applied Computing*, Melbourne, FL, USA, 2003, p. 922–926.
- CHEN, C.; YU, Y. Empirical studies of information visualization: a meta-analysis. *International Journal on Human-Computer Studies*, v. 53, n. 5, p. 851–866, 2000.
- CHENG, J.; KE, Y.; NG, W. delta-tolerance closed frequent itemsets. In: *Proceedings of IEEE International Conference on Data Mining*, Washington, DC, USA: IEEE Computer Society, 2006, p. 139–148.
- CHENG, J.; KE, Y.; NG, W. Effective elimination of redundant association rules. *Data Mining and Knowledge Discovery*, v. 16, n. 2, p. 221–249, 2008.
- CHERNOFF, H. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, v. 68, p. 361–368, 1973.
- CLEVELAND, W. S. *Visualizing data*. Hobart Press, 1993.
- COUTURIER, O.; HAMROUNI, T.; YAHIA, S. B.; NGUIFO, E. M. A scalable association rule visualization towards displaying large amounts of knowledge. In: *Proceedings of International Conference on Information Visualization*, Zurich, Swiss, 2007a, p. 657–663.
- COUTURIER, O.; ROUILLARD, J.; CHEVRIN, V. An interactive approach to display large sets of association rules. In: *Human-Computer Interaction*, Beijing, China, 2007b, p. 258–267.
- DOMINGUES, M. A. *Generalização de regras de associação*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Univ. de São Paulo, 2004.
- DOMINGUES, M. A.; REZENDE, S. O. Using taxonomies to facilitate the analysis of association rules. In: *International Workshop on Knowledge Discovery and Ontologies held within European Conference Principles and Practice of Knowledge Discovery in Databases*, Porto, Portugal, 2005, p. 59–66.
- ERTEK, G.; DEMIRIZ, A. A framework for visualizing association mining results. In: *Proceedings of International Symposium on Computer and Information Sciences*, Istanbul, Turkey, 2006, p. 593–602.

- FALOUTSOS, C.; LIN, K.-I. Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In: *Proceedings of ACM International Conference on Management of Data*, 1995, p. 163–174.
- FAYYAD, U. M.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, n. 3, p. 37–54, 1996a.
- FAYYAD, U. M.; PIATETSKY-SHAPIO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, v. 39, n. 11, p. 27–34, 1996b.
- FEINER, S.; BESHES, C. Metaphors for exploring n-dimensional virtual worlds. In: *Proceedings of ACM Symposium on User Interface Software and Technology*, Snowbird, UT, USA, 1990, p. 76–83.
- FURNAS, G. W.; BUJA, A. Prosections Views: Dimensional Inference through Sections and Projections. *Journal of Computational and Graphical Statistics*, v. 3, n. 4, p. 323–353, 1994.
- GOETHALS, B.; BUSSCHE, J. V. A priori versus a posteriori filtering of association rules. In: *Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, Philadelphia, PA, USA, 1999, p. 1055–1061.
- GOETHALS, B.; BUSSCHE, J. V. On supporting interactive association rule mining. In: *Proceedings of International Conference on Data Warehousing and Knowledge Discovery*, London, UK: Springer-Verlag, 2000, p. 307–316.
- HEARST, M. Tilebars: Visualization of term distribution information in full text information access. In: *Proceedings of ACM Conference on Human Factors in Computing Systems*, Denver, CO, USA, 1995, p. 59–66.
- HIDBER, C. Online association rule mining. In: *Proceedings of ACM International Conference on Management of Data*, Philadelphia, PA, USA, 1999, p. 145–156.
- HOFFMAN, P. E. *Table visualizations: A formal model and its applications*. Tese de Doutorado, Computer Science Dept., Univ. of Massachusetts at Lowell, 1999.
- HOFMANN, H.; SIEBES, A. P. J. M.; WILHELM, A. F. X. Visualizing association rules with interactive mosaic plots. In: *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA, 2000, p. 227–235.

- INSELBERG, A. Multidimensional detective. In: *Proceedings of IEEE Symposium on Information Visualization*, Phoenix, AZ, USA, 1997, p. 100–107.
- INSELBERG, A.; DIMSDALE, B. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In: *Proceedings of IEEE Visualization*, San Francisco, CA, USA, 1990, p. 361–378.
- IVKOVIC, S.; YEARWOOD, J.; STRANIERI, A. Visualizing association rules for feedback with the legal system. In: *Proceedings of International Conference on Artificial Intelligence and Law*, Edinburgh, Scotland, UK, 2003, p. 214–223.
- JOHNSON, B.; SHNEIDERMAN, B. Treemaps: A space-filling approach to the visualization of hierarchical information structures. In: *Proceedings of IEEE Visualization*, San Diego, CA, USA, 1991, p. 284–291.
- JORGE, A. Hierarchical clustering for thematic browsing and summarization of large sets of association rules. In: *Proceedings of SIAM International Conference on Data Mining*, Lake Buena Vista, FL, USA, 2004.
- KEIM, D. A. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, v. 6, n. 1, p. 59–78, 2000.
- KEIM, D. A. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, v. 8, n. 1, p. 1–8, 2002.
- KEIM, D. A.; ANKERST, M.; KRIEGEL, H.-P. Recursive pattern: A technique for visualizing very large amounts of data. In: *Proceedings of IEEE Visualization*, Atlanta, GA, USA, 1995, p. 279–286.
- KEIM, D. A.; HAO, M. C.; DAYAL, U. Hierarchical pixel bar charts. *IEEE Transactions on Visualization and Computer Graphics*, v. 8, n. 3, p. 255–269, 2002.
- KEIM, D. A.; HAO, M. C.; LADISCH, J.; HSU, M.; DAYAL, U. Pixel bar charts: A new visualization technique for visualizing large multi-attribute data sets without aggregation. In: *Proceedings of IEEE Symposium on Information Visualization*, San Diego, CA, USA, 2001, p. 113–120.
- KEIM, D. A.; KRIEGEL, H.-P. Visdb: Database exploration using multidimensional visualization. *IEEE Computer Graphics and Applications*, v. 14, n. 5, p. 40–49, 1994.

- KEIM, D. A.; KRIEGER, H.-P. Visualization techniques for mining large databases: A comparison. *IEEE Transactions on Knowledge and Data Engineering*, v. 8, n. 6, p. 923–938, 1996.
- KLEMETTINEN, M.; MANNILA, H.; RONKAINEN, P.; TOIVONEN, H.; VERKAMO, A. I. Finding interesting rules from large sets of discovered association rules. In: *Proceedings of ACM International Conference on Information and Knowledge Management*, Gaithersburg, MD, USA, 1994, p. 401–407.
- KOPANAKIS, I.; THEODOULIDIS, B. Visual data mining modeling techniques for the visualization of mining outcomes. *Journal of Visual Languages & Computing*, v. 14, n. 6, p. 543–589, 2003.
- KRYSZKIEWICZ, M. Representative association rules and minimum condition maximum consequence association rules. In: *Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery*, London, UK: Springer-Verlag, 1998, p. 361–369.
- KUNTZ, P.; GUILLET, F.; LEHN, R.; BRIAND, H. A user-driven process for mining association rules. In: *Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery*, Lyon, France, 2000, p. 160–168.
- LAMPING, J.; RAO, R.; PIROLI, P. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In: *Proceedings of ACM International Conference on Human Factors in Computing Systems*, Denver, CO, USA, 1995, p. 401–408.
- LEBLANC, J.; WARD, M. O.; WITTELS, N. Exploring n-dimensional databases. In: *Proceedings of IEEE Visualization*, San Francisco, CA, USA, 1990, p. 230–237.
- LEVKOWITZ, H. Color icons: Merging color and texture perception for integrated visualization of multiple parameters. In: *Proceedings of IEEE Visualization*, San Diego, CA, USA, 1991, p. 164–170.
- LEVKOWITZ, H. *Color Theory and Modeling for Computer Graphics, Visualization, and Multimedia Applications*. 1st edition ed Kluwer Academic Publishers, p. 109–133, 1997.
- LEVKOWITZ, H.; PICKETT, R. M.; SMITH, S.; TORPEY, M. An environment and studies for exploring auditory representations of multidimensional data. *Perceptual Issues in Visualization*, p. 47–58, 1995.

- LI, G.; HAMILTON, H. J. Basic association rules. In: *Proceedings of SIAM International Conference on Data Mining*, 2004.
- LIU, Y.; SALVENDY, G. Design and evaluation of visualization support to facilitate association rules modeling. *International Journal of Human-Computer Studies*, v. 21, n. 1, p. 15–38, 2006.
- LÔBO, R.; BEZERRA, L.; FARIA, C.; MAGNABOSCO, C.; ALBUQUERQUE, L.; BERGMANN, J.; SAINZ, R.; OLIVEIRA, H. *Avaliação genética de touros e matrizes da raça nelore: Sumário*. Associação Nacional de Criadores e Perquisadores, 2008.
- MACKINLAY, J. D.; ROBERTSON, G. G.; CARD, S. K. The perspective wall: Detail and context smoothly integrated. In: *Proceedings of ACM International Conference on Human Factors in Computing Systems*, New Orleans, LA, USA, 1991, p. 173–176.
- MAHANTI, A.; ALHAJJ, R. Visual interface for online watching of frequent itemset generation in apriori and eclat. In: *Proceedings of International Conference on Machine Learning and Applications*, Los Angeles, CA, USA, 2005, p. 404–410.
- MELANDA, E. A.; REZENDE, S. O. Combining quality measures to identify interesting association rules. In: LEMAÎTRE, C.; REYES, C. A.; GONZÁLEZ, J. A., eds. *Proceedings of Ibero-American Conference on Artificial Intelligence*, Puebla, México, 2004, p. 441–453.
- MINGHIM, R.; PAULOVICH, F. V.; LOPES, A. A. Content-based text mapping using multi-dimensional projections for exploration of document collections. In: ERBACHER, R. F.; ROBERTS, J. C.; GRÖHN, M. T.; BORNER, K., eds. *Proceedings of SPIE-IS&T Electronic Imaging - Visualization and Data Analysis - VDA*, San Jose, California: SPIE Press, 2006, p. 60600S.
- NOMELINI, J.; REZENDE, S. O.; YAMAMOTO, C. H.; BEZERRA, L. A. F.; VOZZI, P. A.; LÔBO, R. B. Uso de regras de associação para identificação de forças de mercado que regem o comércio de touros com avaliação genética. *Revista Brasileira de Zootecnia*, 2009.
- OLIVEIRA, M. C. F.; LEVKOWITZ, H. From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics*, v. 9, n. 3, p. 378–394, 2003.
- ONG, K.-H.; ONG, K.-L.; NG, W.-K.; LIM, E.-P. Crystalclear: Active visualization of association rules. In: *Proceedings of International Workshop on Active Mining, in conj. IEEE International Conference on Data Mining*, Maebashi, Japan, 2002.

- PALSHIKAR, G. K.; KALE, M. S.; APTE, M. M. Association rules mining using heavy itemsets. *Data Knowledge and Engineering*, v. 61, n. 1, p. 93–113, 2007.
- PASQUIER, N.; TAOUIL, R.; BASTIDE, Y.; STUMME, G.; LAKHAL, L. Generating a condensed representation for association rules. *Journal of Intelligent Information Systems*, v. 24, n. 1, p. 29–60, 2005.
- PAULOVICH, F. V. *Mapeamento de dados multi-dimensionais – integrando mineração e visualização*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, 2008.
- PAULOVICH, F. V.; MINGHIM, R. Text map explorer: a tool to create and explore document maps. In: *Proceedings of International Conference on Information Visualization*, London - UK: IEEE CS Press, 2006, p. 245–251.
- PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least square projection: a fast high precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, v. 14, n. 3, p. 564–575, 2008.
- PAULOVICH, F. V.; OLIVEIRA, M. C. F.; MINGHIM, R. The projection explorer: A flexible tool for projection-based multidimensional visualization. In: *Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing - SIBGRAPI*, Belo Horizonte, Brazil: IEEE CS Press, 2007, p. 27–36.
- PICKETT, R. M.; GRINSTEIN, G. G. Iconographic displays for visualizing multidimensional data. In: *Proceedings of IEEE Conference on Systems, Man and Cybernetics*, 1988, p. 514–519.
- PLAISANT, C. The challenge of information visualization evaluation. In: *Proceedings of Conference on Advanced Visual Interfaces*, New York, NY, USA: ACM, 2004, p. 109–116.
- REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. F. *Sistemas inteligentes: Fundamentos e aplicações* Editora Manole, p. 307–335, 2003.
- RIUL JR., A.; SANTOS JR., D. S. D.; WOHNATH, K.; TOMMAZO, R. D.; CARVALHO, A. C. P. L. F.; FONSECA, F. J.; OLIVEIRA JR., O. N.; TAYLOR, D. M.; MATTOSO, L. H. C. Artificial taste sensor efficient combination of sensors made from langmuir-blodgett films of conducting polymers and a ruthenium complex and

- self-assembled films of an azobenzene-containing polymer. *Langmuir*, v. 18, n. 1, p. 239–245, 2002.
- ROBERTSON, G. G.; MACKINLAY, J. D.; CARD, S. K. Cone Trees: Animated 3D visualizations of hierarchical information. In: *Proceedings of ACM International Conference on Human Factors in Computing Systems*, New Orleans, LA, USA, 1991, p. 189–194.
- SARKAR, M.; BROWN, M. H. Graphical fisheye views. *Communications of the ACM*, v. 37, n. 12, p. 73–84, 1994.
- SHNEIDERMAN, B. Tree visualization with Treemaps: 2-d space-filling approach. *ACM Transactions on Graphics*, v. 11, n. 1, p. 92–99, 1992.
- SHNEIDERMAN, B. Dynamic queries for visual information seeking. *IEEE Software*, v. 11, n. 6, p. 70–77, 1994.
- SHNEIDERMAN, B. The eyes have it: A task by data type taxonomy for information visualizations. In: *Proceedings of IEEE Symposium on Visual Languages*, Boulder, CO, USA, 1996, p. 336–343.
- SPENCE, R.; TWEEDIE, L.; DAWKES, H.; SU, H. Visualization for functional design. In: *Proceedings of IEEE International Symposium on Information Visualization*, Atlanta, Georgia, USA, 1995, p. 4–10.
- SRIKANT, R.; AGRAWAL, R. Mining generalized association rules. In: *Proceedings of International Conference on Very Large Data Bases*, Zurich, Switzerland, 1995, p. 407–419.
- TAN, P.-N.; KUMAR, V.; SRIVASTAVA, J. Selecting the right objective measure for association analysis. *Information Systems*, v. 29, n. 4, p. 293–313, 2004.
- TECHAPICHETVANICH, K.; DATTA, A. Visar: A new technique for visualizing mined association rules. In: *Proceedings of International Conference on Advanced Data Mining and Applications*, Wuhan, China, 2005, p. 88–95.
- TEJADA, E.; MINGHIM, R.; NONATO, L. G. On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization*, v. 2, n. 4, p. 218–231, 2003.

- UNWIN, A.; HOFMANN, H.; BERNT, K. The twokey plot for multiple association rules control. In: *Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery*, 2001.
- WARD, M. O. Xmdvtool: Integrating multiple methods for visualizing multivariate data. In: *Proceedings of IEEE Visualization*, Washington, DC, USA, 1994, p. 326–333.
- WHITE, R. W.; KULES, B.; DRUCKER, S. M.; SCHRAEFEL, M. C. Supporting exploratory search: Introduction. *Communications of the ACM*, v. 49, n. 4, p. 36–39, 2006.
- WIJK, J. J.; LIERE, R. D. Hyperslice. In: *Proceedings of IEEE Conference on Visualization*, 1993, p. 119–125.
- WONG, P. C.; WHITNEY, P.; THOMAS, J. Visualizing association rules for text mining. In: *Proceedings of IEEE Symposium on Information Visualization*, San Francisco, CA, USA, 1999, p. 120–123.
- YAMAMOTO, C. H. Manual do sistema I₂E. 2008.
- YAMAMOTO, C. H.; OLIVEIRA, M. C. F. Visualization to assist users in association rule mining tasks. In: *Workshop de Teses e Dissertações em Banco de Dados, em conjunto com o Simpósio Brasileiro de Banco de Dados*, Uberlândia, MG, 2005.
- YAMAMOTO, C. H.; OLIVEIRA, M. C. F.; FUJIMOTO, M. L.; REZENDE, S. O. An itemset-driven cluster-oriented approach to extract compact and meaningful sets of association rules. In: *Proceedings of International Conference on Machine Learning and Applications*, Washington, DC, USA: IEEE Computer Society, 2007, p. 87–92.
- YAMAMOTO, C. H.; OLIVEIRA, M. C. F.; REZENDE, S. O. Visualization to assist the generation and exploration of association rules. In: ZHAO, Y.; ZHANG, C.; CAO, L., eds. *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*, IGI Global, p. 226–247, 2009.
- YAMAMOTO, C. H.; OLIVEIRA, M. C. F.; REZENDE, S. O.; NOMELINI, J. Including the user in the knowledge discovery loop: Interactive itemset-driven rule extraction. In: *Proceedings of ACM Symposium on Applied Computing*, 2008, p. 1212–1217.
- YANG, L. Visualizing frequent itemsets, association rules and sequential patterns in parallel coordinates. In: *Proceedings of International Conference on Computational Science and Its Applications*, London, UK, 2003, p. 21–30.

- ZAKI, M. J. Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, v. 9, n. 3, p. 223–248, 2004.
- ZAKI, M. J.; PHOOPHAKDEE, B. *Mirage: A framework for mining, exploring and visualizing minimal association rules*. Relatório Técnico, Rensselaer Polytechnic Institute, Computer Sciences Department, 2003.

Arquitetura do Sistema I₂E

A arquitetura do sistema I₂E é mostrada na **Figura A.1**. Os principais módulos que constituem a arquitetura são: *Preprocessor* (pré-processador), *Lib* (biblioteca) e *System* (sistema). Este último, por sua vez, é dividido em seis submódulos: *Apriori*, *Projector* (projektor), *k-means*, *RuleGen*, *I₂E UI* e *RulEx UI*. Cada um dos módulos é descrito a seguir.

Figura A.1: Arquitetura do sistema I₂E.

- **Preprocessor** – Contém classes para pré-processamento de dados para todas as bases utilizadas neste projeto. Estas classes contêm, de maneira geral, rotinas para

leitura, seleção, codificação e gravação de arquivos no formato padrão do módulo I₂E.

- **Lib** – Contém classes para *jitter* (perturbação na posição de objetos na visualização, de modo a evitar sobreposição de objetos), *ranking* (ordenação de regras) e formatação de textos e tratamento de arquivos.
- **System** – Contém as classes principais para chamada dos módulos I₂E e RulEx e a tela de *splash*, além dos seguintes pacotes.
 - **Apriori** – Classes que correspondem ao núcleo da implementação do algoritmo para obtenção de *itemsets* frequentes, incluindo rotinas para construção da estrutura de dados e para extração de *itemsets*.
 - **Projector** – Classes para gerar projeções, utilizando a abordagem *IDMAP*, descrita na **Seção 2.3**. Algumas classes foram fornecidas por Fernando Paulovich (Paulovich et al., 2007).
 - **k-means** – Contém classes para particionamento do conjunto de *itemsets*, utilizando o algoritmo *k*-médias, que é simples e eficiente. Algumas classes foram fornecidas por Fernando Paulovich (Paulovich et al., 2007).
 - **RuleGen** – Classes que correspondem ao núcleo da implementação do algoritmo para geração de regras de associação, incluindo rotinas para seleção de regras representativas.
 - **I₂E UI (User Interface)** – Contém classes para a criação de toda a interface de usuário, incluindo os controles de todas as janelas (painel de controle, lista de itens, visualização de *itemsets* frequentes, lista de regras e controle de log), a barra de ferramentas e os menus/submenus.
 - **RulEx UI (User Interface)** – Contém classes para a criação de toda a interface de usuário, incluindo os controles de todas as janelas (listas de regras, visualização de regras de associação) e os menus. Contém também classes para criação da visualização de medidas de interesse de regras de associação utilizando coordenadas paralelas.

Publicações relacionadas à tese

- (Yamamoto e de Oliveira, 2005) – Yamamoto, C. H.; Oliveira, M. C. F. de “Visualization to assist users in association rule mining tasks”. Apresentado no *IV Workshop de Teses e Dissertações em Banco de Dados (WTDBD)*, Uberlândia-MG, Brasil, 2005.
- (Yamamoto et al., 2007) – Yamamoto, C. H.; Oliveira, M. C. F. de; Fujimoto, M. L.; Rezende, S. O. “An itemset-driven cluster-oriented approach to extract compact and meaningful sets of association rules”. *Anais do IEEE International Conference on Machine Learning and Applications (ICMLA’07)*, Cincinnati-OH, USA, 2007, pp 87–92, <http://dx.doi.org/10.1109/ICMLA.2007.17>.
- (Yamamoto et al., 2008) – Yamamoto, C. H.; Oliveira, M. C. F. de; Rezende, S. O.; Nomelini, J. “Including the User in the Knowledge Discovery Loop: Interactive Itemset-Driven Rule Extraction”. *Anais do ACM Symposium on Applied Computing (SAC’08)*, Multimedia and Visualization Track (MMV), Fortaleza-CE, Brasil, 2008, pp. 1212–1217, <http://doi.acm.org/10.1145/1363686.1363967>.
- (Yamamoto et al., 2009) – Yamamoto, C. H.; Oliveira, M. C. F. de; Rezende, S. O. “Visualization to Assist the Generation and Exploration of Association Rules”. Em Y. Zhao, C. Zhang e L. Cao (Eds.) *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*, 2009, pp. 226–247, IGI Global, Detalhes em www.igi-global.com/reference/details.asp?id=33406.

- (Nomelini et al., 2009) – Nomelini, J.; Rezende, S. O.; Yamamoto, C. H.; Bezerra, L. A. F.; Vozzi, P. A.; Lôbo, R. B. “Uso de Regras de Associação para Identificação de Forças de Mercado que Regem o Comércio de Touros com Avaliação Genética”. (artigo em preparação para a *Revista Brasileira de Zootecnia*, Detalhes em www.rbz.ufv.br/rbz).
- (Yamamoto, 2008) – Yamamoto, C. H. “I₂E System User’s Manual”. Disponibilizado em <http://infoserver.lcad.icmc.usp.br/infovis2/I2E>. Setembro, 2008.