Uma metodologia para exploração de regras de associação
generalizadas integrando técnicas de visualização de
informação com medidas de avaliação do conhecimento

Magaly Lika Fujimoto

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: 24/06/2008

Assinatura:_____

Uma metodologia para exploração de regras de associação generalizadas integrando técnicas de visualização de informação com medidas de avaliação do conhecimento

Magaly Lika Fujimoto

Orientadora: *Prof^a Dr^a Solange Oliveira Rezende*

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

USP – São Carlos
Junho/2008

Este documento foi preparado utilizando-se o formatador de textos L^AT_EX. Sua bibliografia é gerada automaticamente pelo B_IB_TE_X, utilizando o estilo Chicago.

©Copyright 2008 – Magaly Lika Fujimoto
Todos os Direitos Reservados

*À minha família,
à minha orientadora Solange Oliveira Rezende,
e ao meu namorado Cláudio Haruo Yamamoto.*

Agradecimentos

Aos meus pais, Nobuko e Shiguemitsu, pelo apoio incondicional, carinho e educação. Eles sempre prezaram o meu estudo, fazendo todo o possível para me proporcionar uma boa educação e formação. Foi graças ao incentivo deles que eu cheguei até aqui.

À professora Solange, que me orientou desde a iniciação científica no fim do meu primeiro ano de graduação em 2003. Muito obrigada pelos anos de ensinamento, não apenas no âmbito científico/acadêmico, mas também no pessoal.

Ao meu namorado Cláudio, pelo seu amor e paciência. Agradeço também pelas correções, por sanar minhas dúvidas de visualização e por me acalmar nos momentos de “pânico”.

Aos meus amigos da graduação, Giselle e Paulo, que me apoiaram na época de estresse da trilha.

À Roberta, Veronica e Edson Melanda que me ajudaram desde a época da iniciação. Também agradeço a Veronica por me ajudar na definição do escopo deste projeto de mestrado.

À Fernanda, Merley e Fabiano pelo apoio no dia anterior à entrega desta dissertação. Também agradeço a Fernanda pelos conselhos que ajudaram a definir o título deste projeto.

Ao Anand pelas correções de inglês.

À todos os alunos e professores do LABIC pela agradável convivência nesses anos.

Aos funcionários do ICMC pela dedicação, competência e respeito.

À Universidade de São Paulo pela oportunidade e pela qualidade de ensino e pesquisa.

Ao CNPq que permitiu a realização deste trabalho.

A todos que, direta ou indiretamente, contribuíram para a realização desta dissertação.

Resumo

O processo de mineração de dados tem como objetivo encontrar o conhecimento implícito em um conjunto de dados para auxiliar a tomada de decisão. Do ponto de vista do usuário, vários problemas podem ser encontrados durante a etapa de pós-processamento e disponibilização do conhecimento extraído, como a enorme quantidade de padrões gerados por alguns algoritmos de extração e a dificuldade na compreensão dos modelos extraídos dos dados. Além do problema da quantidade de regras, os algoritmos tradicionais de regras de associação podem levar à descoberta de conhecimento muito específico. Assim, pode ser realizada a generalização das regras de associação com o intuito de obter um conhecimento mais geral. Neste projeto é proposta uma metodologia interativa que auxilie na avaliação de regras de associação generalizadas, visando melhorar a compreensibilidade e facilitar a identificação de conhecimento interessante. Este auxílio é realizado por meio do uso de técnicas de visualização em conjunto com a aplicação medidas de avaliação objetivas e subjetivas, que estão implementadas no módulo de visualização de regras de associação generalizados denominado RULEE-GARVIS, que está integrado ao ambiente de exploração de regras RULEE (*Rule Exploration Environment*). O ambiente RULEE está sendo desenvolvido no LABIC-ICMC-USP e auxilia a etapa de pós-processamento e disponibilização de conhecimento. Neste contexto, também foi objetivo deste projeto de pesquisa desenvolver o Módulo de Gerenciamento do ambiente de exploração de regras RULEE. Com a realização do estudo dirigido, foi possível verificar que a metodologia proposta realmente facilita a compreensão e a identificação de regras de associação generalizadas interessantes.

Abstract

The data mining process aims at finding implicit knowledge in a data set to aid in a decision-making process. From the user's point of view, several problems can be found at the stage of post-processing and provision of the extracted knowledge, such as the huge number of patterns generated by some of the extraction algorithms and the difficulty in understanding the types of the extracted data. Besides the problem of the number of rules, the traditional algorithms of association rules may lead to the discovery of very specific knowledge. Thus, the generalization of association rules can be realized to obtain a more general knowledge. In this project an interactive methodology is proposed to aid in the evaluation of generalized association rules in order to improve the understanding and to facilitate the identification of interesting knowledge. This aid is accomplished through the use of visualization techniques along with the application of objective and subjective evaluation measures, which are implemented in the visualization module of generalized association rules called RULEE-GARVIS, which is integrated with the Rule Exploration Environment RULEE. The RULEE environment is being developed at LABIC-ICMC-USP and aids in the post-processing and provision of knowledge. In this context, it was also the objective of this research project to develop the Module Management of the rule exploration environment RULEE. Through this directed study, it was verified that the proposed methodology really facilitates the understanding and identification of interesting generalized association rules.

Sumário

Agradecimentos	ix
Resumo	xi
Abstract	xiii
Sumário	xv
Lista de Figuras	xix
1 Introdução	1
1.1 Hipótese	4
1.2 Objetivos	4
1.3 Organização	5
2 Mineração de Dados e Associação	7
2.1 Considerações Iniciais	7
2.2 O Processo de Mineração de Dados	8
2.2.1 Identificação do Problema	9
2.2.2 Pré-processamento	10
2.2.3 Extração de Padrões	12
2.2.4 Pós-processamento	13
2.2.5 Utilização do Conhecimento	14
2.2.6 Considerações sobre Mineração de Dados	15
2.3 Associação	15
2.3.1 Conceitos e Definições	16
2.3.2 Regras de Associação Generalizadas	18
2.4 Considerações Finais	21

3	Abordagens para Avaliação de Regras	23
3.1	Considerações Iniciais	23
3.2	Medidas de Avaliação de Conhecimento	25
3.2.1	Medidas Objetivas	25
3.2.2	Medidas Subjetivas	29
3.3	Técnicas de Visualização de Informação para Apoiar a Compreensibilidade de Regras	34
3.3.1	Classificação das Técnicas de Visualização de Informação	34
3.3.2	Visualização de Regras de Associação	37
3.4	Trabalhos Relacionados com Avaliação de Regras	44
3.4.1	O aplicativo AIAS	44
3.4.2	O ambiente PEAR	45
3.4.3	Módulo de visualização de regras de associação de Chakravarthy e Zhang	47
3.4.4	O sistema VisAR	48
3.4.5	O sistema I ₂ E e RulEx	50
3.4.6	O ambiente RULEE	52
3.5	Considerações Finais	59
4	Metodologia para Incrementar a Compreensibilidade de Regras de Associação Generalizadas e Apoiar a Identificação de Conhecimento Interessante	61
4.1	Considerações Iniciais	61
4.2	Metodologia Proposta	63
4.3	Considerações sobre a Metodologia Proposta	67
4.4	Módulo de Visualização de Regras de Associação Generalizadas - RULEE-GARVIS	69
4.4.1	Requisitos para a utilização do Módulo RULEE-GARVIS	69
4.4.2	Funcionalidades do Módulo RULEE-GARVIS	70
4.5	Estudo dirigido da aplicação da metodologia utilizando o módulo RULEE-GARVIS	75
4.6	Considerações Finais	80
5	Módulo de Gerenciamento do Ambiente RuleE	83
5.1	Considerações Iniciais	83
5.2	Detalhamento do Módulo de Gerenciamento	84
5.3	Considerações Finais	93

6 Conclusão e Trabalhos Futuros	95
Referências Bibliográficas	101

Lista de Figuras

2.1	Etapas do processo de mineração de dados	9
2.2	Exemplo de uma taxonomia para vestuário	19
3.1	Classificação das técnicas de visualização de informação (Keim & Ward, 2003)	35
3.2	<i>Mosaic Plot</i> e a sua derivação <i>Double Decker Plots</i> (Hofmann, Siebes, & Wilhelm, 2000)	38
3.3	<i>Enhanced Grid View</i> e <i>Tree View</i> implementadas no CrystalClear (Ong, Ong, Ng, & Lim, 2002)	40
3.4	Metáfora visual de Blanchard, Guillet, & Briand (2003)	40
3.5	Abordagem utilizada por Chakravarthy & Zhang (2003)	41
3.6	Exemplos do modelo de Bruzzese & Buono (2004)	41
3.7	Técnica VisAR (Techapichetvanich & Datta, 2005)	42
3.8	Exemplos de visualização de <i>itemsets</i> freqüentes e regras de associação em coordenadas paralelas (Yang, 2005)	43
3.9	Visualização com coordenadas paralelas Yang (2005)	43
3.10	Aplicativo AIAS (Liu, Hsu, Chen, & Ma, 2000)	46
3.11	O ambiente PEAR (Jorge, Poças, & Azevedo, 2002)	46
3.12	Visualização das medidas suporte e confiança em gráficos X-Y e histograma (Jorge, Poças, & Azevedo, 2002)	47
3.13	Visualização de regras de associação no formato de tabelas (Chakravarthy & Zhang, 2003)	48
3.14	Visualização 2D e 3D (Chakravarthy & Zhang, 2003)	48
3.15	Visualização de regras de associação no VisAR (Techapichetvanich & Datta, 2005)	49
3.16	O sistema I ₂ E (Yamamoto, Oliveira, & Rezende, 2008)	50

3.17	O sistema RuleX (Yamamoto, Oliveira, & Rezende, 2008)	51
3.18	Arquitetura geral do ambiente RULEE	53
3.19	Metodologia para identificação de regras de associação interessantes (Sinoara, 2006)	57
4.1	Metodologia que integra técnicas de visualização de informação com medidas objetivas e subjetivas	64
4.2	Diagrama de Caso de Uso do RULEE-GARVIS	71
4.3	Selecionar subconjunto de Regras Potencialmente Interessantes	77
4.4	Selecionar subconjunto RPI - <i>Taxonomia</i>	78
4.5	Avaliar Regras Potencialmente Interessantes - Avaliar subconjunto RPI . .	78
4.6	Avaliar Regras Potencialmente Interessantes - Exemplo de gráfico disponível	79
4.7	Iniciar exploração com medidas subjetivas	79
4.8	Visualização das Informações de uma Exploração Encerrada	80
5.1	Diagrama de casos de uso do núcleo básico do RULEE	85
5.2	Esquema do núcleo básico do RULEE	86
5.3	Diagrama de casos de uso do módulo RULEE-RAG	87
5.4	Esquema do módulo RULEE-RAG	87
5.5	Diagrama de casos de uso do módulo RULEE-SEAR - contido em Sinoara, Fujimoto, & Rezende (2006)	88
5.6	Esquema do módulo RULEE-SEAR	88
5.7	Diagrama de casos de uso do módulo RULEE-SACT	89
5.8	Esquema do módulo RULEE-SACT	90
5.9	Diagrama de casos de uso do módulo <i>ARInE</i>	90
5.10	Esquema da interface para os novos módulos	91
5.11	Código do protótipo do ambiente RULEE	92
5.12	Código com a consulta SQL substituída pela chamada da Biblioteca de Acesso	92

Introdução

A evolução da computação promoveu o desenvolvimento das tecnologias de coleta e de armazenamento de dados, que permitiram um enorme acúmulo de informações armazenadas em grandes bases de dados (Kurgan & Musilek, 2006). No caso de organizações, essas informações podem trazer vantagens competitivas, além da possibilidade de conhecer melhor os seus clientes. Assim, as organizações têm investido na aquisição e desenvolvimento de ferramentas de análise de dados, uma vez que os métodos manuais tornaram-se dispendiosos, subjetivos e inviáveis, quando aplicados a grandes bases de dados. Diante da deficiência dos métodos manuais, diversas pesquisas têm sido direcionadas ao desenvolvimento de tecnologias de extração automática de conhecimento a partir de dados. Esse campo de pesquisa é chamado de extração de conhecimento de base de dados ou mineração de dados (MD).

O processo de mineração de dados tem o objetivo de encontrar conhecimento a partir de um conjunto de dados para ser utilizado em um processo de tomada de decisão. Para apoiar as etapas do processo de mineração de dados, está sendo desenvolvido o ambiente DISCOVER no Laboratório de Inteligência Computacional (LABIC¹) do Insti-

¹<http://labic.icmc.usp.br>.

tuto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP²). O ambiente DISCOVER foi proposto com o objetivo de fornecer um ambiente integrado para apoiar as etapas do processo de extração de conhecimento de dados e textos. Nesse ambiente são utilizados algoritmos de Aprendizado de Máquina implementados pela comunidade científica, bem como módulos com finalidades específicas desenvolvidos pelos pesquisadores do LABIC. Entre as funcionalidades desses módulos estão: pré-processamento de dados e de textos e pós-processamento de conhecimento.

Na linha do pós-processamento do conhecimento, está sendo desenvolvido no LABIC, no âmbito do DISCOVER, um ambiente para exploração e disponibilização de regras, denominado RULEE (*Rule Exploration Environment*). O RULEE permite a exploração de regras de classificação, regressão e associação. O ambiente RULEE está sendo desenvolvido devido à necessidade de participação de usuários especialistas do domínio na exploração do conhecimento visando a identificação de conhecimento interessante.

Do ponto de vista do usuário, um dos problemas encontrados no final do processo de mineração de dados é que muitos dos algoritmos de extração geram uma enorme quantidade de padrões (Melanda & Rezende, 2004). Por exemplo, algoritmos de extração de regras de associação geralmente produzem milhares ou dezenas de milhares de regras, especialmente quando os atributos da base de dados são altamente correlacionados. Isso dificulta a compreensão dos modelos extraídos e a identificação de um possível conhecimento interessante. Assim, uma maior atenção tem sido direcionada a apoiar os usuários na identificação do conhecimento interessante (Liu, Hsu, Chen, & Ma, 2000; Hilderman & Hamilton, 2001; Tan, Kumar, & Srivastava, 2002; Omiecinski, 2003; Melanda & Rezende, 2004; Tan, Kumar, & Srivastava, 2004; Natarajan & Shekar, 2005; Carvalho, Freitas, & Ebecken, 2005; Sinoara & Rezende, 2006; Tamir & Singer, 2006; Yamamoto, Oliveira, Fujimoto, & Rezende, 2007; Yamamoto, Oliveira, & Rezende, 2008).

Uma forma para amenizar o problema do grande número de regras de associação é tornar o conhecimento específico mais geral. Isto pode ser realizado desde que exista algum conhecimento sobre o domínio da aplicação, que pode ser representado via taxonomias.

²<http://www.icmc.usp.br>.

Apesar das regras de associação generalizadas permitirem a exploração do conhecimento em diferentes níveis de abstração, ainda existe a necessidade de encontrar uma forma de explorar a potencialidade dessas regras. Assim, este trabalho tem como um dos seus objetivos contribuir com uma metodologia para auxiliar na compreensão e na identificação de regras de associação generalizadas interessantes. Para tal foram utilizadas técnicas de visualização de informação com apoio de medidas de avaliação objetivas e subjetivas. Desta forma, foram aproveitadas as vantagens oferecidas pelo uso em conjunto das medidas objetivas e subjetivas, já investigadas por Sinoara (2006), além das vantagens que as técnicas de visualização de informação proporcionam, devido a facilidade de interpretação visual das informações por parte dos usuários.

Este trabalho possui como entrada as regras de associação generalizadas utilizando uma taxonomia existente, sendo que a generalização é realizada utilizando o algoritmo *APRA_{alg}* proposto por Carvalho (2007). Também é importante observar, que este trabalho adapta o fluxo de análise proposto por Sinoara (2006). Em Sinoara (2006) é proposto a utilização de medidas objetivas e subjetivas para avaliação de regras de associação, sendo que a análise objetiva é realizada no módulo *ARInE* (Melanda, 2004) e a análise subjetiva no RULEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006), a serem descritos na Seção 3.4.6. Ressaltando que a metodologia de Sinoara (2006), assim como os módulos *ARInE* e RULEE-SEAR, foram idealizados para avaliação de regras de associação.

Assim, este trabalho possui como diferencial a incorporação de visualização interativa, de aspectos particulares de regras de associação generalizadas e a possibilidade de análise de regras de complemento, exceção e redundantes, a serem descritos na Seção 4.2. Para viabilizar a metodologia proposta neste trabalho, foi desenvolvido um módulo denominado RULEE-GARVIS, no qual é realizado tanto a análise objetiva como a subjetiva. Um dos principais motivos que levaram o módulo *ARInE* não ser utilizado, foi que todos os seus gráficos são estáticos, não havendo a possibilidade de interação direta com nenhum deles. Já com relação ao módulo RULEE-SEAR, foi o fato de não disponibilizar gráficos para análise de regras. Além disso, como citado, os módulos *ARInE* e RULEE-SEAR foram desenvolvidos para avaliação de regras de associação, não considerando os aspectos

particulares de regras de associação generalizadas.

O módulo RULEE-GARVIS desenvolvido neste trabalho é integrante do ambiente RULEE. Apesar do ambiente RULEE disponibilizar diversas ferramentas para exploração de regras, não existia até então uma ferramenta de exploração de regras de associação generalizadas que realmente auxiliasse o usuário. O RULEE possuía apenas o módulo RULEE-RAG (Fujimoto, Carvalho, & Rezende, 2007) para exploração de regras de associação generalizadas, que permite a exploração textual das regras. Devido a essa limitação de exploração, o módulo RULEE-RAG não auxiliava adequadamente o usuário na compreensão e identificação de conhecimentos interessantes.

Visando aproveitar algumas funcionalidades oferecidas pelo ambiente RULEE, como inserção de regras e cálculo de medidas, também é objetivo deste trabalho, desenvolver o Módulo de Gerenciamento para apoiar o desenvolvimento do ambiente de exploração de regras RULEE. O seu desenvolvimento permite a implementação do módulo RULEE-GARVIS segundo a arquitetura do ambiente RULEE.

1.1 Hipótese

A hipótese é que técnicas de visualização de informação com apoio de medidas de avaliação objetivas e subjetivas, facilitam a compreensão e a identificação do conhecimento interessante em regras de associação generalizadas. Essa combinação aproveita as vantagens que cada técnica e medida proporcionam.

1.2 Objetivos

O objetivo principal deste trabalho é auxiliar a compreensibilidade e facilitar a identificação de conhecimento interessante na análise de regras de associação generalizadas, por meio da utilização de técnicas de visualização de informação em conjunto com medidas objetivas e subjetivas. Para tal, propor uma metodologia e implementar um módulo que viabilize avaliar a sua eficácia.

Também é objetivo deste projeto de pesquisa desenvolver o Módulo de Gerenciamento do ambiente para exploração de regras RULEE baseado na arquitetura descrita na Seção 3.4.6. O ambiente deve ser implementado de forma a ampliar as funcionalidades e suprir as deficiências existentes no protótipo atual. Para alcançar estas metas, o ambiente RULEE deve utilizar a Base de Dados e a Biblioteca de Acesso que foram modeladas e implementadas no projeto de iniciação científica financiado pela FAPESP³. Desse modo, em relação ao desenvolvimento do ambiente RULEE, o objetivo deste projeto está no desenvolvimento do Módulo de Gerenciamento e a sua integração aos módulos já desenvolvidos no projeto de iniciação científica.

É importante ressaltar que apesar do projeto constituir-se de dois objetivos, o Módulo de Gerenciamento viabiliza a implementação do módulo de visualização segundo a nova arquitetura do ambiente RULEE. Desse modo, o módulo de visualização poderá beneficiar-se de algumas funcionalidades oferecidas pelo ambiente RULEE.

1.3 Organização

Visando atingir os objetivos descritos na seção anterior e confirmar a hipótese, a seguir é apresentada a organização desta dissertação. No Capítulo 2 são apresentados o processo de mineração de dados, a tarefa de associação e as regras de associação generalizadas. No Capítulo 3 são apresentados alguns métodos e técnicas usados para avaliação de regras, como medidas para avaliação do conhecimento e técnicas de visualização de informação. Também são apresentados alguns trabalhos relacionados com avaliação de regras. No Capítulo 4 é apresentada a metodologia para auxiliar na compreensão e identificação de regras de associação generalizadas interessantes, bem como o módulo RULEE-GARVIS e um estudo dirigido da aplicação da metodologia utilizando o módulo RULEE-GARVIS. No Capítulo 5 é descrito o desenvolvimento do Módulo de Gerenciamento. Por fim são apresentadas as conclusões e alguns trabalhos futuros no Capítulo 6.

³Processo 05/52244-4 - Modelagem e Implementação da Base de Dados e da Biblioteca de Acesso do Ambiente RULEE- Magaly Lika Fujimoto - 01/08/2005 a 31/07/2006.

Mineração de Dados e Associação

2.1 *Considerações Iniciais*

O desenvolvimento de tecnologias de armazenamento de dados promoveu o crescimento da quantidade de dados disponíveis nas bases de dados das organizações. Desta forma, a utilização de técnicas e ferramentas tradicionais em grandes bases de dados tornaram-se dispendiosas. Para suprir essa deficiência, diversas pesquisas têm sido direcionadas ao desenvolvimento de tecnologias de extração automática de conhecimento a partir de dados. Esse campo de pesquisa é chamado de extração de conhecimento de base de dados ou mineração de dados (MD) (Rezende, 2004).

A mineração de dados foi definida por Fayyad, Piatetsky-Shapiro, & Smyth (1996) como sendo: “O processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados”. Desse modo, neste capítulo estão descritos o processo de mineração de dados, as suas etapas e uma de suas tarefas, a associação. A associação é caracterizada pela extração de regras de associação. Utilizando um conhe-

cimento de domínio no processo de mineração de dados, é possível realizar a descoberta de regras de associação generalizadas, que representam conhecimento mais geral. O conhecimento de domínio pode ser representado hierarquicamente, com níveis variados de abstração de conceitos, organizados por um especialista de domínio ou um processo automatizado, denominado taxonomia. Assim, neste capítulo também são apresentadas as regras de associação generalizadas.

2.2 *O Processo de Mineração de Dados*

Existem diversas abordagens para a divisão das etapas do processo de mineração de dados. Inicialmente, foi proposto em Fayyad, Piatetsky-Shapiro, & Smyth (1996) uma divisão do processo em nove etapas. Já em Weiss & Indurkha (1998), essa divisão é composta por apenas quatro etapas. Entretanto, neste trabalho é considerada a divisão do processo em um ciclo composto de cinco grandes etapas: identificação do problema, pré-processamento dos dados, extração de padrões, pós-processamento do conhecimento e utilização do conhecimento obtido (Rezende, Pugliesi, Melanda, & Paula, 2003). Essas etapas são ilustradas na Figura 2.1 e descritas nas seções seguintes.

Observa-se que, normalmente, esse processo é iterativo e interativo, pois não se pode esperar que a extração de um conhecimento útil seja realizado simplesmente submetendo um conjunto de dados a uma “caixa preta” (Mannila, 1996). Essa iteratividade e interatividade do processo de mineração de dados é centrado na interação entre os usuários, que podem ser divididos em três classes: **Especialista do Domínio**, usuário que deve possuir amplo conhecimento do domínio da aplicação e fornecer apoio à execução do processo; **Analista**, usuário especialista no processo de mineração de dados e responsável por sua execução; e **Usuário Final**, usuário que utiliza o conhecimento extraído no processo de mineração de dados para apoiar a tomada de decisão. Não é necessário que este usuário possua conhecimento profundo do domínio da aplicação.

Figura 2.1: Etapas do processo de mineração de dados (Rezende, Pugliesi, Melanda, & Paula, 2003)

2.2.1 *Identificação do Problema*

Nesta etapa são realizados estudos para identificar o problema e adquirir um conhecimento inicial do domínio. As restrições, os objetivos e as metas a serem alcançadas no processo de mineração de dados são definidos. Também são identificados e selecionados os conjuntos de dados a serem utilizados para a extração de conhecimento.

O conhecimento sobre o domínio fornece um subsídio para todas as etapas do processo de mineração de dados (Rezende, Pugliesi, Melanda, & Paula, 2003). Na etapa de pré-processamento, esse conhecimento pode ajudar os analistas na escolha do melhor conjunto de dados para realizar a extração de padrões, dos critérios de preferência entre os possíveis atributos, das restrições de relacionamento ou das informações para geração de novos atributos. Na etapa de extração de padrões, este pode ajudar os analistas na escolha de um critério de preferência entre os modelos gerados, ou mesmo na geração de um

conhecimento inicial a ser fornecido como entrada do algoritmo de mineração para melhorar a precisão ou a compreensão do modelo final. Já na etapa de pós-processamento, o conhecimento extraído pelos algoritmos de extração de padrões deve ser avaliado e alguns critérios de avaliação utilizam o conhecimento do especialista para saber, por exemplo, se o conhecimento extraído é interessante ao usuário.

A definição dos objetivos também é uma atividade muito importante, uma vez que os objetivos definidos guiarão o processo de mineração de dados.

2.2.2 *Pré-processamento*

Após a identificação do problema, com o entendimento do domínio da aplicação e considerando aspectos como os objetivos e as fontes de dados (bases de dados das quais se pretende extrair o conhecimento), inicia-se a etapa de pré-processamento. É uma etapa que pode ser fortemente (tarefas realizadas somente com uso de conhecimento específico de domínio) ou fracamente (tarefas podem ser realizadas por métodos que extraem dos próprios dados as informações necessárias para tratar o problema) dependente de conhecimento de domínio (Batista, 2003). Nesta etapa são realizados a seleção e o tratamento dos dados a partir dessas fontes, de acordo com os objetivos identificados para o processo de mineração de dados. Geralmente o tratamento nos dados é necessário devido aos dados selecionados não estarem em um formato adequado para a extração de conhecimento. Além disso, durante o processo de coleta de dados podem ocorrer diversos problemas, como erros de digitação e geração de dados incorretos ou inconsistentes, por exemplo. As atividades que podem ser realizadas com a finalidade de tratar esses dados estão descritas a seguir.

Extração e Integração - os dados disponíveis podem estar em diferentes formatos, como arquivos-texto, arquivos no formato de planilhas, banco de dados ou *data warehouse*. Assim, é necessário a obtenção e a unificação desses dados, formando uma única fonte (Han & Kamber, 2006).

Transformação - algumas transformações que podem ser aplicadas aos dados são (Batista,

2003): a normalização de atributos contínuos, que é realizada para colocar os valores em intervalos definidos como, por exemplo, entre 0 e 1; a discretização de atributos quantitativos, como transformar esses atributos em faixas de valores (qualitativos); a transformação de atributos qualitativos em quantitativos, por exemplo, atributos qualitativos com ordem podem ser mapeados de forma numérica mantendo essa ordem, como pequeno = 1, médio = 2, grande = 3; e a transformação de tipo como, converter um atributo do tipo data em um outro tipo aceito pelo algoritmo.

Limpeza - para garantir a qualidade dos dados é necessária a aplicação de algumas técnicas de limpeza, pois os dados podem apresentar problemas resultantes do processo de coleta, como erros de digitação ou leitura dos dados por sensores. A limpeza também pode ser aplicada a dados que não interessam ao processo.

Redução de Dados - a aplicação de métodos para redução de dados pode ser muito útil quando o número de exemplos e de atributos disponíveis para análise torna inviável a utilização de algoritmos de extração de padrões. Essa redução pode ser feita de três modos (Weiss & Indurkha, 1998): reduzindo o número de exemplos por meio da geração de amostras representativas dos dados (Glymour, Madigan, Pregibon, & Smyth, 1997), a fim de manter as características do conjunto de dados original; reduzindo o número de atributos para diminuir o espaço de busca pela solução, porém mantendo a qualidade final da solução (para tal é aconselhável o apoio do especialista do domínio, pois a remoção de um atributo potencialmente útil pode diminuir a qualidade do conhecimento extraído); e reduzindo o número de valores de um atributo, aplicando métodos como a discretização, que é a substituição de um atributo contínuo por um atributo discreto por meio do agrupamento de seus valores, ou a suavização de valores de um atributo contínuo, que agrupa os valores de um determinado atributo e o substitui por um valor numérico que o represente.

É importante salientar que esse tratamento deve ser realizado com o devido cuidado nas atividades descritas, uma vez que é fundamental garantir que as informações presentes nos dados originais continuem presentes nas amostras geradas. Assim, os modelos finais irão

representar o comportamento dos dados originais. Ao final da etapa de pré-processamento os dados estão prontos para serem submetidos a um algoritmo da etapa de extração de padrões. Esta etapa está descrita na próxima seção.

2.2.3 *Extração de Padrões*

Na etapa de extração de padrões é realizada a escolha da tarefa de mineração de dados a ser empregada, a escolha do algoritmo a ser utilizado e a extração dos padrões propriamente dita. O objetivo é encontrar padrões/modelos (conhecimento) a partir dos dados. Portanto, a escolha da tarefa é muito importante e deve ser realizada de acordo com os objetivos desejáveis para a solução a ser encontrada, pois sua escolha determina o tipo do conhecimento extraído. As possíveis tarefas de mineração de dados podem ser agrupadas em atividades preditivas e descritivas.

A mineração de dados preditiva consiste na generalização de exemplos ou experiências passadas com respostas conhecidas em uma linguagem capaz de identificar a classe (atributo meta) de um novo exemplo. As duas principais tarefas na predição são a classificação e a regressão. A classificação consiste na predição de um valor categórico, por exemplo, prever se o cliente é bom ou mau pagador. Na regressão, o atributo a ser predito consiste em um valor contínuo, por exemplo, prever o lucro ou a perda em um empréstimo. Já a mineração de dados descritiva consiste na identificação de comportamentos intrínsecos do conjunto de dados, sendo que estes dados não possuem uma classe especificada. Algumas das tarefas de descrição são *clustering*, sumarização e associação.

Existe uma grande variedade de algoritmos que podem ser utilizados para executar cada tarefa. A escolha do algoritmo é feita de maneira subordinada à linguagem de representação dos padrões a serem encontrados. Entre os tipos mais frequentes de representação de padrões encontram-se: árvores de decisão, regras de produção, modelos lineares, modelos não-lineares (redes neurais artificiais), modelos baseados em exemplos (*k-nearest neighbour*, raciocínio baseado em casos) e modelos de dependência probabilística (redes bayesianas). Segundo Kohavi, Sommerfield, & Dougherty (1996) não existe

um algoritmo mais adequado para todas as tarefas de mineração de dados. Assim, pode-se utilizar diferentes algoritmos durante a etapa de extração de padrões, gerando vários modelos que serão tratados na etapa de pós-processamento a fim de selecionar um bom modelo para o usuário final.

A extração dos padrões propriamente dita consiste na configuração dos parâmetros e na aplicação dos algoritmos selecionados para extrair os padrões contidos nos dados. É importante ressaltar que, dependendo do problema e dos objetivos, podem ser necessárias diversas execuções dos algoritmos. Por exemplo, para obter um classificador mais preciso, pode ser necessária a combinação de vários outros (Rezende, 2004). Já no caso da tarefa de associação, descrita na Seção 2.3, pode-se realizar apenas uma execução do algoritmo sem comprometer a qualidade do conhecimento extraído. Segundo Zheng, Kohavi, & Mason (2001), dada a mesma entrada os algoritmos para obter regras de associação devem gerar a mesma solução. Além disso, a alteração dos valores de suporte e confiança mínimos, parâmetros de entrada normalmente utilizados durante a geração de regras, apenas incluirá ou excluirá regras do conjunto de regras extraídas.

A disponibilização do conjunto de padrões extraídos nesta etapa ou a sua incorporação a um sistema inteligente ocorre após a análise e/ou o processamento dos padrões, realizados na etapa de pós-processamento.

2.2.4 Pós-processamento

Após a etapa de extração de padrões, deve ser realizada a etapa de pós-processamento, na qual o conhecimento extraído pode ser simplificado, avaliado, visualizado ou simplesmente documentado para o usuário final. Essa etapa consiste de vários métodos e procedimentos que podem ser agrupados nas categorias apresentadas a seguir (Bruha & Famili, 2000).

Filtragem do Conhecimento - pode ser realizada empregando a restrição de atributos ou a ordenação de regras por meio de métricas (medidas de avaliação), por exemplo.

Interpretação e Explicação - usualmente aplicada quando o conhecimento obtido é utilizado por um usuário final ou por um sistema inteligente, podendo ser documentado, visualizado ou modificado de forma a torná-lo compreensível ao usuário. O conhecimento extraído pode ser comparado ao preexistente para a verificação de conflitos ou de conformidade, podendo ser sumarizado e/ou combinado com o conhecimento prévio do domínio.

Avaliação - pode ser realizada verificando a precisão, a compreensão, a complexidade computacional, o interesse, entre outros.

Integração do Conhecimento - os sistemas tradicionais de apoio à decisão são dependentes de uma única técnica, estratégia e modelo. Já os sistemas novos e sofisticados possibilitam combinar ou refinar os resultados de vários modelos de maneira a obter uma maior precisão e um melhor desempenho.

Analisando o conhecimento extraído pode-se determinar, por exemplo, se o processo de extração deve ser repetido ou não. Caso o conhecimento extraído não seja interessante ao usuário ou não esteja de acordo com os objetivos pré-estabelecidos, pode ser necessária a realização de etapas específicas do processo de mineração de dados, ou de todo o processo, ajustando-se os parâmetros utilizados ou realizando-se melhorias na seleção de dados, entre outros. Caso contrário, o conhecimento pode ser disponibilizado ao usuário final para o uso na fase de utilização do conhecimento, descrita na próxima seção.

2.2.5 Utilização do Conhecimento

Depois de ser avaliado e validado no pós-processamento, o conhecimento é consolidado nesta etapa. Este pode ser utilizado para resolver eventuais conflitos entre o conhecimento prévio e o obtido pelo processo de mineração de dados (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Além disso, o conhecimento pode ser incorporado a um sistema inteligente, utilizado diretamente pelo usuário final para apoio a algum processo de tomada de decisão ou, simplesmente, relatado às pessoas interessadas.

2.2.6 Considerações sobre Mineração de Dados

As empresas têm uma grande quantidade de dados armazenados que podem possuir informações valiosas, como tendências e padrões que podem ser usados, por exemplo, para tornar as decisões de negócios mais eficientes. Assim, tornou-se necessário o desenvolvimento de processos de análise automática, como o processo de mineração de dados. Contudo, existe um aspecto importante sobre esse processo a ser considerado. Quando a mineração de dados é aplicada a problemas reais, a etapa de pós-processamento e a disponibilização do conhecimento obtido tornam-se decisivas para o sucesso do processo. A avaliação desse conhecimento obtido é importante para se garantir a qualidade e a precisão dos modelos.

Além da avaliação, a disponibilização do conhecimento também é importante, visto que o processo de mineração de dados é interativo. É necessário apoiar o acesso dos usuários ao conhecimento descoberto, pois a participação destes na identificação de conhecimento interessante durante o pós-processamento é imprescindível. Após essa etapa, o conhecimento também deve ficar disponível aos usuários, para que possa ser utilizado diretamente em processos de tomada de decisão ou em sistemas inteligentes.

Na próxima seção é apresentada uma tarefa de mineração de dados que tem grande aplicabilidade a problemas reais, a associação, caracterizada pela extração de regras de associação.

2.3 Associação

A associação é uma tarefa de mineração de dados que tem despertado grande interesse (Baesens, Viaene, & Vanthienen, 2000), tanto na área acadêmica como em aplicações práticas. Pesquisas vêm sendo desenvolvidas e organizações têm utilizado seus resultados em aplicações relacionadas a comércio, contratos de seguro, geoprocessamento, serviços bancários, medicina, entre outros (Antonie & Zaiane, 2004; Aggelis, 2004; Besemann, Denton, Yekkirala, Hutchison, & Anderson, 2004; Oliveira, Zaane, & Saygin, 2004; Ribeiro,

Balan, Felipe, Traina, & Traina, 2005; Kumar, Yip, Smith, & Grenon, 2006; Maalouf & Mansour, 2007; Ribeiro, Traina, Traina, & Azevedo-Marques, 2008).

2.3.1 Conceitos e Definições

Associação é uma tarefa de mineração de dados classificada como uma atividade descritiva. Essa tarefa visa descobrir o quanto um conjunto de itens presentes em um registro de uma base de dados implica na presença de algum outro conjunto distinto de itens no mesmo registro (Agrawal & Srikant, 1994). Assim, com a extração de regras de associação é possível encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados.

O formato de uma regra de associação pode ser representado como uma implicação na forma $LHS \Rightarrow RHS$, em que LHS e RHS são, respectivamente, o lado esquerdo (*Left Hand Side*) e o lado direito (*Right Hand Side*) da regra, definidos por conjuntos disjuntos de itens. As regras de associação podem ser definidas como descrito a seguir (Agrawal & Srikant, 1994).

Seja D uma base de dados composta por um conjunto de itens $A = \{a_1, \dots, a_m\}$ ordenados lexicograficamente e por um conjunto de transações $T = \{t_1, \dots, t_n\}$, na qual cada transação $t_i \in T$ é composta por um conjunto de itens (*itemset*), tal que $t_i \subseteq A$.

A regra de associação é uma implicação na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$ e $LHS \cap RHS = \emptyset$. A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações T com confiança $conf$ se em $conf\%$ das transações de T em que LHS ocorre, RHS também ocorre. A regra $LHS \Rightarrow RHS$ tem suporte sup se em $sup\%$ das transações em T ocorrem $LHS \cup RHS$.

Em regras de associação, as medidas mais empregadas são o suporte e a confiança, tanto no que se refere à etapa de pós-processamento do conhecimento adquirido, como na etapa de seleção dos subconjuntos de itens durante o processo de geração de regras.

Suporte - representa a probabilidade de ocorrência de um *itemset* X ou da transação $LHS \cup RHS$ no conjunto de dados. Da maneira como foi definido, o suporte para um *itemset* X pode ser representado por:

$$sup(X) = P(X) \quad (2.1)$$

Já o suporte de uma regra $LHS \Rightarrow RHS$ pode ser representado por:

$$sup(LHS \Rightarrow RHS) = sup(LHS \cup RHS) = P(LHS \cup RHS) \quad (2.2)$$

Confiança - indica a frequência com que LHS e RHS ocorrem juntos em relação ao número total de transações em que LHS ocorre, ou seja, probabilidade condicional de RHS dado LHS . Do modo como foi definida, a confiança de uma regra $LHS \Rightarrow RHS$ pode ser representada por:

$$conf(LHS \Rightarrow RHS) = P(RHS/LHS) \quad (2.3)$$

Em outras palavras, o suporte representa as frequências dos padrões e a confiança a força da implicação, ou seja, em pelo menos $c\%$ das vezes que o antecedente ocorrer nas transações, o conseqüente também deve ocorrer (Zhang & Zhang, 2002).

O problema de obtenção de regras de associação é decomposto em dois sub-problemas (Agrawal, Imielinski, & Swami, 1993):

1. Encontrar todos os k -*itemsets* (conjunto de k itens) que possuam suporte maior ou igual ao suporte mínimo especificado pelo usuário. Os *itemsets* com suporte igual ou superior ao suporte mínimo especificado são definidos como *itemsets* frequentes, os demais conjuntos são denominados de *itemsets* não-frequentes;
2. Utilizar todos os k -*itemsets* frequentes, com $k \geq 2$, para gerar as regras de associação. Para cada *itemset* frequente $l \subseteq A$, encontrar todos os subconjuntos $\tilde{a}$ de itens não vazios de l . Para cada subconjunto $\tilde{a} \subseteq l$, gerar uma regra na forma

$\tilde{a} \Rightarrow (l - \tilde{a})$ se a razão de $\text{sup}(l)$ por $\text{sup}(\tilde{a})$ é maior ou igual a confiança mínima especificada pelo usuário.

Com um conjunto de *itemsets* freqüentes $\{a, b, c, d\}$ e um subconjunto de *itemsets* freqüentes $\{a, b\}$, por exemplo, pode-se gerar uma regra do tipo $ab \Rightarrow cd$, desde que $\text{conf}(ab \Rightarrow cd) \geq$ confiança mínima especificada, em que, $\text{conf}(ab \Rightarrow cd) = \text{sup}(a, b, c, d) / \text{sup}(a, b)$.

Como descrito, devem ser definidos valores mínimos de suporte e confiança antes de se realizar a extração de regras de associação. Um problema ao se definir esses valores é que, geralmente, se eles forem altos são geradas regras triviais e, se forem baixos são gerados um grande volume de regras, dificultando a análise por parte do usuário. Uma forma de auxiliar o usuário na análise de um grande volume de regras é generalizar as regras, ou seja, tornar mais gerais os conceitos específicos, expressando um conhecimento mais amplo da realidade e facilitando a sua compreensão, sem perder as regras específicas. As regras de associação generalizadas são discutidas na seção seguinte.

2.3.2 Regras de Associação Generalizadas

A generalização de regras de associação torna mais gerais os conceitos específicos. Para que a generalização de regras ocorra, é necessário algum conhecimento sobre o domínio da aplicação, podendo ser expresso, por exemplo, via taxonomias. As taxonomias refletem uma caracterização coletiva ou individual de como os itens podem ser hierarquicamente classificados (Adamo, 2001). Eventualmente, múltiplas taxonomias podem estar presentes simultaneamente, refletindo a existência de diversos pontos de vista ou a possibilidade de classificações distintas para o mesmo conjunto de itens. Na Figura 2.2 é apresentado um pequeno exemplo de uma taxonomia. Nesse exemplo pode-se verificar que: camiseta é uma roupa leve, bermuda é uma roupa leve, roupa leve é um tipo de roupa, sandália é um tipo de calçado, etc.

Por exemplo, considerar que existem as regras “quem compra camiseta também compra tênis” ($\text{Camisetas} \Rightarrow \text{Tênis}$), e “quem compra bermuda também compra tênis” (Bermudas

Figura 2.2: Exemplo de uma taxonomia para vestuário

$\Rightarrow$ Tênis). Com a utilização da taxonomia da Figura 2.2, é possível generalizar essas regras e dizer que “quem compra roupas leves também compra tênis” ($\text{Roupas Leves} \Rightarrow \text{Tênis}$).

A seguir são apresentados alguns conceitos necessários para se ter uma melhor compreensão do uso de taxonomias em regras de associação. Estes conceitos foram definidos em Srikant & Agrawal (1997) e Adamo (2001).

Generalização ou Ancestral - considerando X um *itemset*, em que $X \subseteq LHS$ ou $X \subseteq RHS$ e $LHS \Rightarrow RHS$ uma regra de associação, as notações $X\uparrow$ e $(LHS \Rightarrow RHS)\uparrow$ representam novos *itemsets* e regras que, respectivamente, derivam de X e $LHS \Rightarrow RHS$ pela substituição de um ou mais itens pelos seus ancestrais na taxonomia. Os novos *itemsets* e regras são ditas generalizações ou ancestrais de X e $LHS \Rightarrow RHS$.

Especialização ou Descendente - considerando X um *itemsets*, em que $X \subseteq LHS$ ou $X \subseteq RHS$ e $LHS \Rightarrow RHS$ uma regra de associação, as notações $X\downarrow$ e $(LHS \Rightarrow RHS)\downarrow$ representam novos *itemset* e regras que, respectivamente, derivam de X e $LHS \Rightarrow RHS$ pela substituição de um ou mais itens pelos seus descendentes na taxonomia. Os novos *itemsets* e regras são ditas especializações ou descendentes de X e $LHS \Rightarrow RHS$.

Pai Um *itemset* $X\uparrow$ é dito ser pai de X se não há nenhum *itemset* X' tal que X' é um ancestral de X e $X\uparrow$ é um ancestral de X' .

Filho Um *itemset* $X\downarrow$ é dito ser filho de X se não há nenhum *itemset* X' tal que X' é um descendente de X e $X\downarrow$ é um descendente de X' .

Generalização máxima Um *itemset* X é dito ser uma generalização máxima, se nenhum item em X pode ser substituído por um item ancestral na taxonomia. O mesmo é válido para uma regra $LHS \Rightarrow RHS$ e para um item a .

Especialização máxima Um *itemset* X é dito ser uma especialização máxima, se nenhum item em X pode ser substituído por um item descendente na taxonomia. O mesmo também é válido para uma regra $LHS \Rightarrow RHS$ e para um item a .

Assim, uma regra de associação generalizada usando taxonomias pode ser definida como (Srikant & Agrawal, 1997):

Seja D uma base de dados composta por um conjunto de itens $A = \{a_1, \dots, a_m\}$ ordenados lexicograficamente e por um conjunto de transações $T = \{t_1, \dots, t_n\}$, na qual cada transação $t_i \in T$ é composta por um conjunto de itens tal que $t_i \subseteq A$. É dito que uma transação t_i suporta um item $a_j \in A$, se a_j está em t_i ou a_j é um ancestral de algum item em t_i . Seja $\mathcal{T}$ um grafo direcional e acíclico com os itens, representando um conjunto de taxonomias. Se há uma aresta em $\mathcal{T}$ de um item $a_p \in A$ para um item $a_c \in A$, a_p é dito ser pai de a_c e a_c é dito ser filho de a_p .

Uma regra de associação generalizada usando taxonomia é uma implicação na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$, $LHS \cap RHS = \emptyset$ e nenhum item em RHS é um ancestral de qualquer item em LHS . A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações T com confiança $conf$ se em $conf\%$ das transações de T em que ocorre LHS ocorre também RHS . A regra $LHS \Rightarrow RHS$ tem suporte sup se em $sup\%$ das transações de T ocorre $LHS \cup RHS$.

É importante salientar que nas regras de associação generalizadas, o cálculo do suporte não é a soma do suporte dos itens filhos. Assim, o suporte para um item a_j não terminal na taxonomia, pode ser definido como:

$$sup(a_j) = \frac{\# desc(a_j)}{N},$$

em que $desc(a_j)$ é o conjunto de descendentes mais próximos de a_j , sendo que $desc(a_j) = \bigcup desc(a_z)$, tal que a_z seja um descendente mais próximos de a_j . Na definição do suporte para um item não terminal, o símbolo “#” indica a cardinalidade de um conjunto e N é o número total de transações consideradas.

Na literatura existem muitos trabalhos que utilizam taxonomias em regras de associação, diferindo apenas na etapa do processo de mineração de dados em que as taxonomias são aplicadas. As taxonomias podem ser utilizadas nas etapas de pré-processamento, extração de padrões ou pós-processamento. Segundo Carvalho (2007), das possibilidades de aplicação da taxonomia, a mais interessante é que as taxonomias sejam utilizadas na etapa de pós-processamento, uma vez que a utilização de conhecimento de fundo pode melhorar a análise dos padrões obtidos. Este trabalho utiliza a abordagem proposta por Carvalho (2007), que generaliza as regras de associação na etapa de pós-processamento do conhecimento utilizando o algoritmo denominado $APRA_{alg}$. Apesar da generalização poder melhorar a análise dos padrões, ainda há o problema de explorar a potencialidade das regras de associação generalizadas, de modo a permitir uma exploração em que o usuário possua diversas ferramentas para compreender melhor o conjunto e identificar regras interessantes. Além disso, realizar uma exploração que permita um fácil acesso às regras específicas se for o desejo do usuário.

2.4 Considerações Finais

Nesse capítulo foi apresentado o processo de mineração de dados, que tem grande utilidade na aquisição automática de conhecimento a partir de grandes bases de dados.

Além disso, foi fornecida uma visão geral sobre as regras de associação e as regras de associação generalizadas. Para tanto, foram apresentadas suas definições e conceitos.

Um dos problemas de regras de associação é o grande volume de conhecimento (no formato de regras) que é gerado, podendo dificultar a análise e a compreensão do conhecimento pelo usuário. Assim, generalizar regras utilizando taxonomias é uma forma de amenizar este problema, pois reduz o volume das regras geradas. Entretanto, apenas generalizar as regras não é garantia de que as regras resultantes serão compreensíveis e interessantes ao usuário. Assim, ainda há a necessidade de aplicar métodos e técnicas para avaliá-las, com o intuito de melhorar a compreensão e identificar regras interessantes. Desse modo, algumas abordagens usadas para a avaliação de regras são descritas no próximo capítulo.

Abordagens para Avaliação de Regras

3.1 *Considerações Iniciais*

A avaliação de regras é uma das principais atividades realizadas durante a etapa de pós-processamento do processo de mineração de dados, podendo ser verificada a precisão, a compreensão, a complexidade computacional, o interesse, entre outros. Na avaliação da qualidade das regras são utilizados conceitos como compreensibilidade e interessabilidade. A compreensibilidade está relacionada à facilidade de interpretação das regras por parte dos usuários e a interessabilidade refere-se ao grau de interesse de uma determinada regra para um usuário e está relacionada a fatores como novidade, utilidade, relevância e significância estatística.

Diversas técnicas têm sido utilizadas para a avaliação das regras obtidas na extração de padrões do processo de mineração de dados e, conseqüentemente, auxiliado na identificação do conhecimento de interesse. Algumas dessas técnicas são a avaliação por consulta, a aplicação de medidas de avaliação de conhecimento, a generalização, a poda de regras e

o agrupamento de regras (Melanda, 2004). Essas técnicas são descritas a seguir.

Avaliação por Consulta - faz uso das linguagens de consulta, como o SQL (*Structured Query Language*), para que o usuário explore o conjunto de regras.

Medidas de Avaliação de Conhecimento - fornece subsídios ao usuário com relação ao entendimento e à utilização do conhecimento adquirido. Essas medidas podem ser categorizadas quanto ao modo e objeto de avaliação. Em relação ao modo de avaliação, as medidas podem ser objetivas ou subjetivas. As medidas objetivas dependem exclusivamente da estrutura dos padrões (regras) e dos dados utilizados no processo de mineração de dados. Já as medidas subjetivas dependem fundamentalmente dos usuários que irão interpretar o conhecimento (Silberschatz & Tuzhilin, 1996). Com relação ao objeto de avaliação, as medidas permitem avaliar o desempenho ou a qualidade de uma regra. O desempenho de uma regra está associado à fidelidade com que representa os dados. Para a avaliação da qualidade dos padrões gerados são utilizados os conceitos de compreensibilidade e grau de interesse. A facilidade de um ser humano interpretar um dado conjunto de regras está relacionada à compreensibilidade deste conjunto, podendo ser estimada, por exemplo, pelo número de regras presentes no conjunto e pelo número de condições em cada regra. O grau de interesse, por sua vez, é uma avaliação de natureza qualitativa realizada a partir de estimativas da quantidade de conhecimento interessante (inovador, inesperado) presente nas regras.

Poda de Regras - reduz o número de regras geradas com o objetivo de excluir regras redundantes ou que não são interessantes ao usuário.

Generalização - utiliza taxonomias para transformar regras específicas em conceitos gerais, produzindo conjuntos de regras mais compactos e geralmente mais compreensíveis aos usuários.

Agrupamento de Regras - fornece ao usuário grupos de regras que correspondam a diferentes temas do domínio em questão, procurando identificar agrupamentos que

representam características específicas de determinados grupos de regras contidas no conjunto original. Com base nesse agrupamento, técnicas de generalização podem ser utilizadas para compactar as regras de cada agrupamento encontrado. Cada um desses agrupamentos pode representar particularidades da base de dados.

Além dessas técnicas, a visualização de informação também pode auxiliar na avaliação de regras. Essa avaliação é realizada visualizando as regras projetadas por meio de alguma forma de representação visual. A visualização de informação é interativa e geralmente aproveita a capacidade humana de interpretação visual. Desse modo, este trabalho utilizará técnicas de visualização de informação em conjunto com medidas de avaliação do conhecimento visando aumentar a compreensibilidade e facilitar a identificação do conhecimento de interesse. Considerando o foco deste trabalho, neste capítulo são apresentadas medidas de avaliação objetivas e subjetivas, além de técnicas de visualização de informação visando apoio ao usuário com relação à compreensibilidade das regras. Também são apresentadas algumas formas de visualização de regras de associação e trabalhos relacionados com avaliação de regras.

3.2 Medidas de Avaliação de Conhecimento

Como descrito na seção anterior, as medidas de avaliação de conhecimento auxiliam o usuário no entendimento e na utilização do conhecimento adquirido, sendo que as medidas podem ser objetivas ou subjetivas.

3.2.1 Medidas Objetivas

Considerando o padrão $A \Rightarrow B$, na Tabela 3.1 são apresentadas algumas medidas objetivas utilizadas por Carvalho, Rezende, & Castro (2007), que realizou um estudo das medidas objetivas mais apropriadas para avaliação de regras de associação generalizadas. As medidas apresentadas nesta seção foram selecionadas por possuir um intervalo fechado (bem definido) e a condição de estarem implementadas no ambiente de pós-processamento

denominado RULEE (a ser descrito na Seção 3.4.6). É importante ressaltar que, neste projeto, foram utilizadas medidas com intervalo fechado para as visualizações da metodologia (a ser descrita no Capítulo 4), pelo fato de serem mais facilmente mapeadas para uma representação visual 2D.

As medidas confiança e suporte foram definidas na Seção 2.3.1, pois isso não serão repetidas aqui.

IS/Cosine (IS) A medida IS pode ser interpretada como o cosseno do ângulo entre dois vetores (Tan, Steinbach, & Kumar, 2005). Segundo Tan, Steinbach, & Kumar (2005) a medida IS mede tanto o interesse quanto a significância do padrão.

Jaccard É um coeficiente que mede a similaridade entre conjuntos. Se A e B são dois conjuntos então a similaridade entre eles é medida pela razão entre o número de elementos em comum e o número de elementos diferentes (Lourenço, Lobo, & Bação, 2004). O valor $Jaccard = 1$ implica em uma total correspondência entre os objetos (Borlund & Ingwersen, 1998).

Laplace É uma medida muito utilizada nos algoritmos de indução de regras. A medida Laplace é uma variação da medida confiança e foi desenvolvida com o objetivo de penalizar regras muito específicas, ou seja, regras que cobrem poucos exemplos (transações) para evitar o *overfitting* (Smaldon & Freitas, 2006).

Suponha, por exemplo, que um determinado *itemset* ocorra em apenas uma transação, assim como todos os seus subconjuntos próprios não nulos. Sem a correção proposta pela medida Laplace, a confiança desse *itemset* seria de $100\%(1/1)$ (Equação 3.1), um valor muito otimista para uma regra extremamente específica. Sendo assim, levando-se em consideração o número de classes contidas no domínio pela medida Laplace, obtém-se um valor mais real, que no caso do exemplo acima seria de $67\%((1 + 1)/(1 + 2))$ (Equação 3.5) (Smaldon & Freitas, 2006). O número 2 no denominador indica o número de classes no caso das regras de associação, o qual representa a presença ou ausência de um item em uma transação.

Medidas	Intervalo	Fórmula
Confiança	[0...1]	$Conf = P(B A)$ (3.1)
Suporte	[0...1]	$Sup = P(AB)$ (3.2)
IS/Cosine (IS)	$0... \sqrt{P(AB)}...1$	$IS = \frac{P(AB)}{\sqrt{P(A)P(B)}} = \frac{A \bullet B}{ A \times B } = Cosine$ (3.3)
Jaccard	[0...1]	$\frac{P(A, B)}{P(A) + P(B) - P(A, B)} \equiv \frac{P(A, B)}{P(A \cup B)}$ (3.4)
Laplace	[0...1]	$\frac{N \times P(AB) + 1}{N \times P(A) + 2}$ (3.5)
ϕ -coefficient	[-1...0...1]	$\frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(B)(1 - P(A))(1 - P(B))}}$ (3.6)
Piatetsky-Shapiro's	[-0.25...0...0.25]	$P(A, B) - P(A)P(B)$ (3.7)
Gini	[0...1]	$P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] - P(B)^2 - P(\bar{B})^2$ (3.8)
Added Value	[-1...0...1]	$P(B A) - P(B) = Conf(A, B) - P(B)$ (3.9)
Kappa	[-1...0...1]	$\frac{P(A, B) + P(\bar{A}, \bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$ (3.10)
Certainty Factor	[-1...0...1]	$\frac{P(B A) - P(B)}{1 - P(B)}$ (3.11)

Tabela 3.1: Algumas Medidas Objetivas

ϕ -coefficient Essa medida indica o grau de associação (correlação) entre duas variáveis (Tan, Kumar, & Srivastava, 2004). Um valor de $\phi = -1$ indica uma correlação

negativa perfeita entre A e B , $\phi = 1$ uma correlação positiva perfeita entre A e B e, $\phi = 0$ que não há correlação entre A e B , ou seja, A e B são independentes.

Piatetsky-Shapiro's Também conhecida como *Rule Interest*, *Novelty* e *Leverage*. Essa medida calcula a porcentagem de transações adicionais cobertas por uma regra de associação que estão acima do esperado (Gonçalves & Plastino, 2004). Em outras palavras, compara o valor observado da ocorrência de A e B e o valor esperado de ocorrência se A e B fossem independentes. Se *Piatetsky – Shapiro's* = 0 diz-se que A e B são independentes. Se *Piatetsky – Shapiro's* > 0 diz-se que A e B são dependentes positivamente; caso contrário, dependentes negativamente.

Gini Essa medida é freqüentemente utilizada como medida de seleção de atributo na indução de árvores de decisão. Ela é usada para medir o decréscimo esperado na impureza ou incerteza de uma determinada classe (variável meta), condicionada ao conhecimento do valor de uma determinada variável (variável preditora) (Fisher, 1996). Sendo assim, se duas variáveis estiverem altamente associadas, então a quantidade de redução será grande.

Added Value Essa medida indica o quanto a freqüência do conseqüente aumenta a presença do antecedente, ou seja, mede o ganho de B na presença de A . Se $P(B|A) > P(B)$ tem-se que a freqüência de B aumenta na presença de A . Se $P(B|A) < P(B)$ tem-se que a freqüência de B diminui na presença de A . Se $P(B|A) = P(B)$ tem-se uma coincidência aleatória, ou seja, A não aumenta em nada a freqüência de B (independência estatística). Portanto, quanto maior for o ganho de B em relação a A mais relacionadas estão as variáveis.

Kappa É um coeficiente de concordância. A medida kappa é calculada pela razão $(P(O) - P(E))/(1 - P(E))$, onde $P(O)$ representa a concordância observada entre dois indivíduos, isto é, a proporção de vezes que os indivíduos concordam, e $P(E)$ a concordância esperada, isto é, a proporção de vezes que se espera que os indivíduos concordem por chance (Eugenio & Glass, 2004). Se *Kappa* = 1 tem-se concordância absoluta; se *Kappa* = 0 tem-se que a concordância é igual a chance (isto é, coin-

cidência puramente aleatória); se $Kappa = -1$ tem-se discordância absoluta, isto é, a propensão dos indivíduos em evitar classificações feitas por outros indivíduos. Observação: $P(O) = P(A, B) + P(\bar{A}, \bar{B})$ e $P(E) = P(A)P(B) - P(\bar{A})P(\bar{B})$.

Certainty Factor $P(B)$ reflete a crença em B . Então $1 - P(B)$ pode ser visto como uma estimativa da descrença em relação a verdade de B . Se $P(B|A)$ for maior que $P(B)$, significa que A aumenta a crença em B diminuindo a sua descrença em relação a verdade de B . Sendo assim, essa medida mede o aumento da crença em B em consequência da observação de A . Em outras palavras, mede a diminuição proporcional na descrença da hipótese B como resultado da observação de A . Se $CF = 1$ então $P(B|A) = 1$ (A e B possuem dependência positiva). Se $CF = -1$ então $P(\bar{B}|A) = 1$ (A e B possuem dependência negativa). Se $CF = 0$ significa que A não confirma nem contradiz B , isto é, A e B são independentes.

Nesta seção foram descritas as medidas objetivas que serão utilizadas na visualização de regras de associação generalizadas da metodologia a ser descrita no Capítulo 4. Como argumentado, estas medidas foram selecionadas entre as apresentadas em Carvalho, Rezende, & Castro (2007), que realizou um estudo das medidas objetivas utilizadas para avaliação de regras de associação generalizadas. A seleção destas medidas foram condicionadas à restrição de possuírem um intervalo fechado e estarem implementadas no ambiente de exploração de regras RULEE (descrito na Seção 3.4.6). Uma síntese de outras medidas objetivas são apresentadas em Melanda (2004); Tan, Steinbach, & Kumar (2005); Geng & Hamilton (2006); Pecina & Schlesinger (2006); Carvalho, Rezende, & Castro (2007).

3.2.2 Medidas Subjetivas

Em Silberschatz & Tuzhilin (1996) foi proposta uma classificação para as medidas subjetivas, que considera os aspectos subjetivos (interesse do usuário) na avaliação de regras. Essa classificação identifica as duas principais razões que tornam o conhecimento interessante, que são a inesperabilidade e a utilidade.

Inesperabilidade (*Unexpectedness*) - o conhecimento é interessante se é novo para o usuário ou contradiz seu conhecimento prévio ou sua expectativa.

Utilidade (*Actionability*) - o conhecimento é interessante se o usuário pode tomar alguma decisão com ele obtendo alguma vantagem.

Esses dois conceitos não são excludentes entre si. Regras interessantes podem ser apresentadas de maneira combinada, podendo ser: inesperadas e úteis, inesperadas e não úteis ou esperadas e úteis.

Liu, Hsu, Chen, & Ma (2000) propõem quatro medidas para identificar regras de associação esperadas e inesperadas considerando o conhecimento prévio do domínio. Para captar o conhecimento que o usuário possui sobre o domínio, Liu, Hsu, Chen, & Ma (2000) propõe uma linguagem predefinida, podendo expressar o conhecimento como:

Impressão geral (GI)- relação que o especialista acredita existir entre os itens especificados.

$$gi(< S_1, \dots, S_m >) \text{ [suporte, confiança]}$$

Conhecimento impreciso (RPC)- conhecimento que o especialista supõe ser verdadeiro.

$$rpc(< S_1, \dots, S_m \rightarrow V_1, \dots, V_g >) \text{ [suporte, confiança]}$$

Conhecimento preciso (PK)- o usuário acredita na precisão da associação.

$$pk(< S_1, \dots, S_m \rightarrow V_1, \dots, V_g >) \text{ [suporte, confiança]}$$

Nessa linguagem para especificação do conhecimento, os valores de suporte e confiança são opcionais. O usuário também pode definir uma taxonomia, ou seja, especificar categorias e níveis de abstração para os atributos. As classes definidas nessa taxonomia podem ser utilizadas na especificação dos conhecimentos. Assim, cada elemento S_i (ou

V_i) pode ser um item, uma classe ou uma expressão $C+$ ou $C*$, nas quais C é uma classe. $C+$ e $C*$ correspondem, respectivamente, a uma ou mais, ou zero ou mais, instâncias da classe C .

O conhecimento especificado é utilizado para analisar as regras descobertas. Cada regra descoberta tem sua estrutura comparada com cada conhecimento especificado, a fim de verificar o número de itens que casam com os elementos especificados pelo usuário. A partir dessas análises são calculados os valores das medidas conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados.

As medidas são definidas em relação ao grau com que o LHS e/ou RHS da regra descoberta casa(m) com o conhecimento fornecido pelo usuário. Assim, L_{ij} é um fator que mede o quanto do LHS da regra está em conformidade com o LHS do conhecimento fornecido pelo especialista do domínio. Pode ser considerado o mesmo para R_{ij} em relação a RHS . Os valores destes fatores variam de 0 (nenhuma conformidade) a 1 (conformidade completa).

Conformidade - identifica e classifica regras em conformidade com uma impressão geral ou um conhecimento impreciso fornecido pelo usuário especialista do domínio.

$$confm_{ij} = L_{ij} \cdot R_{ij} \quad (3.12)$$

Antecedente inesperado - avalia se o antecedente (LHS) da regra é inesperado.

$$unexpCond_{ij} = \begin{cases} 0 & \text{se } R_{ij} - L_{ij} \leq 0, \\ R_{ij} - L_{ij} & \text{se } R_{ij} - L_{ij} > 0. \end{cases} \quad (3.13)$$

Conseqüente inesperado - avalia se o conseqüente (RHS) da regra é inesperado.

$$unexpConseq_{ij} = \begin{cases} 0 & \text{se } L_{ij} - R_{ij} \leq 0, \\ L_{ij} - R_{ij} & \text{se } L_{ij} - R_{ij} > 0. \end{cases} \quad (3.14)$$

Antecedente e conseqüente inesperados - avalia se o antecedente e o conseqüente da regra são inesperados.

$$bsUnexp_{ij} = 1 - \max[confm_{ij}, unexpConseq_{ij}, unexpCond_{ij}] \quad (3.15)$$

Os valores de L_{ij} e R_{ij} são calculados de acordo com o tipo de conhecimento fornecido pelo usuário que está sendo utilizado. Se o conhecimento foi fornecido como uma impressão geral, o cálculo de L_{ij} e R_{ij} é dado por:

se

$$\frac{LM_{ij}}{LN_i} > \frac{RM_{ij}}{RN_i}$$

então

$$L_{ij} = \min\left(\frac{LM_{ij}}{LN_i}, \frac{SM_{ij}}{SN_j}\right) \quad (3.16)$$

$$R_{ij} = \frac{RM_{ij}}{RN_i} \quad (3.17)$$

senão

$$R_{ij} = \min\left(\frac{RM_{ij}}{RN_i}, \frac{SM_{ij}}{SN_j}\right) \quad (3.18)$$

$$L_{ij} = \frac{LM_{ij}}{LN_i} \quad (3.19)$$

sendo que,

- LN_i e RN_i representam, respectivamente, o número de itens no antecedente e no conseqüente da regra descoberta;
- SN_j se refere ao número de elementos¹ do conjunto especificado pelo usuário como sendo uma impressão geral. Caso $SN_j = 0$, então a razão $\frac{SM_{ij}}{SN_j} = 1$;
- LM_{ij} e RM_{ij} representam, respectivamente, o número de itens no antecedente e no conseqüente da regra descoberta que casam com os elementos da impressão geral;

¹Conforme foi definido por Liu, Hsu, Chen, & Ma (2000), o conhecimento fornecido pelo usuário pode ser composto por itens e/ou classes de uma taxonomia, aqui chamados de elementos.

- SM_{ij} refere-se ao número de elementos do conjunto especificado pelo usuário que casam com itens da regra descoberta.

Se o conhecimento foi fornecido como um conhecimento impreciso, o cálculo de L_{ij} e R_{ij} é dado pelas equações 3.20 e 3.21:

$$L_{ij} = \min\left(\frac{LM_{ij}}{LN_i}, \frac{LSM_{ij}}{LSN_j}\right) \quad (3.20)$$

$$R_{ij} = \min\left(\frac{RM_{ij}}{RN_i}, \frac{RVM_{ij}}{RVN_j}\right) \quad (3.21)$$

sendo que,

- LN_i e RN_i representam, respectivamente, o número de itens no antecedente e no conseqüente da regra descoberta;
- LSN_j e RVN_j representam, respectivamente, o número de elementos no antecedente e no conseqüente do conhecimento impreciso. Caso $LSN_j = 0$ (ou $RVN_j = 0$) então $\frac{LSM_{ij}}{LSN_j} = 1$ (ou $\frac{RVM_{ij}}{RVN_j} = 1$);
- LM_{ij} e RM_{ij} representam, cada qual, o número de itens no antecedente e no conseqüente da regra descoberta que casam respectivamente com o antecedente e o conseqüente do conhecimento impreciso;
- LSM_{ij} e RVM_{ij} representam, cada qual, o número de elementos no antecedente e no conseqüente do conhecimento impreciso que casam respectivamente com o antecedente e o conseqüente da regra descoberta.

Nesta seção foram apresentadas medidas de avaliação objetivas e subjetivas. Na próxima seção são descritas algumas técnicas de visualização de informação, que também podem ser usadas para auxiliar a avaliação de regras.

3.3 *Técnicas de Visualização de Informação para Apoiar a Compreensibilidade de Regras*

Segundo Card, Mackinlay, & Shneiderman (1999) a visualização de informação é o uso de representação visual, interativa e suportada por computador, de dados abstratos para ampliar a cognição. O objetivo é apoiar a interação entre o usuário e esses dados, facilitando a exploração e aquisição de conhecimentos úteis (Oliveira & Levkowitz, 2003).

As técnicas de visualização de informação podem ser classificadas segundo diferentes critérios. A classificação descrita na próxima seção é a classificação realizada por Keim & Ward (2003).

3.3.1 *Classificação das Técnicas de Visualização de Informação*

As técnicas de visualização de informação podem ser classificadas baseadas em três critérios: tipo de dados a serem visualizados, técnicas de visualização e técnicas de interação e distorção (Keim & Ward, 2003). Essa classificação é exibida na Figura 3.1, mostrando uma ortogonalidade entre esses critérios, significando que qualquer técnica de visualização pode ser usada em conjunto com qualquer técnica de interação/distorção para qualquer tipo de dado.

Segundo Keim & Ward (2003), os tipos de dados a serem visualizados podem ser categorizados como: dados unidimensionais, bidimensionais, multidimensionais, textos e hipertextos, hierarquias e grafos, algoritmos e *softwares*.

Dados unidimensionais - normalmente possuem uma dimensão mais densa. Um exemplo deste tipo de dados são os dados temporais.

Dados bidimensionais - possuem duas dimensões distintas. Um exemplo são os dados geográficos.

Dados multidimensionais - possuem três ou mais dimensões distintas. Um exemplo

Figura 3.1: Classificação das técnicas de visualização de informação (Keim & Ward, 2003)

tradicional são as tabelas de uma base de dados relacional, que geralmente possuem dezenas/centenas de atributos.

Textos e hipertextos - são dados que não são facilmente descritos por números. Por isso, em muitos casos, torna-se necessária a realização de transformações nos dados previamente à aplicação de alguma técnica de visualização. Artigos e documentos *Web* são exemplos deste tipo de dados.

Hierarquias e grafos - são usados para representar hierarquia de conceitos, como o relacionamento entre as pessoas.

Algoritmos e *softwares* - proporcionam suporte ao desenvolvimento de *softwares* por meio do entendimento dos algoritmos. Um exemplo pode ser observado na detecção de erros. A visualização pode melhorar a detecção de erros por facilitar o entendimento do programador.

Para o entendimento destes diferentes tipos de dados, existem diversas técnicas de visualização. As técnicas de visualização descritas por Keim & Ward (2003) são: padrão 2D/3D, projeção geométrica, iconográficas, baseadas em pixels e hierárquicas. A técnica padrão 2D/3D são representadas pelas visualizações em gráficos x-y (bidimensional) ou x-y-z (tridimensional). Um exemplo são os gráficos de barra.

A técnica de projeção geométrica procura encontrar projeções interessantes de dados definidos em espaços multidimensionais, as quais podem ser exibidas em espaços bidimensionais. Um exemplo deste tipo de técnica são as coordenadas paralelas, cuja idéia básica é mapear os dados k -dimensionais em uma visualização bidimensional, desenhando eixos paralelos igualmente espaçados. Cada eixo corresponde a um atributo e cada item é representado por uma poli-linha que intersecta cada eixo em um ponto, que corresponde ao valor associado ao atributo em questão.

A técnica iconográfica procura mapear cada item de dados multidimensionais em um ícone, de forma que as suas características visuais reflitam os valores que os ícones representam. A técnica baseada em pixels, mapeia cada valor de dimensão para um pixel colorido, agrupando os pixels pertencentes a cada dimensão em áreas adjacentes. Já as técnicas hierárquicas subdividem o espaço k -dimensional e apresentam os sub-espacos obtidos de maneira hierárquica.

Dependendo do tipo de dados, a utilização de uma ou a combinação de várias técnicas de visualização em conjunto com técnicas de interação e distorção permite uma melhor interação do usuário com a visualização. Desse modo, possibilita uma melhor exploração dos dados.

As técnicas de interação proporcionam recursos para o usuário interagir com a visualização, mudando-a dinamicamente de acordo com os objetivos da exploração. Já as técnicas de distorção auxiliam no processo de exploração dos dados, possibilitando ao usuário focar em detalhes sem que haja uma perda da visão geral dos dados. Keim & Ward (2003) descreveu as técnicas de projeção dinâmica, filtragem interativa, *zoom* interativo, distorção interativa e ligação e seleção interativas.

Projeção dinâmica - permite mudar dinamicamente as projeções multidimensionais para uma melhor exploração do conjunto de dados. A principal desvantagem desta técnica é o crescimento exponencial do número de projeções à medida do crescimento do número de dimensões.

Filtragem interativa - possibilita selecionar subconjuntos de dados de interesse inter-

ativamente. Esta filtragem pode ser realizada por meio de seleções ou consultas.

Zoom interativo permite que o usuário observe, ao mesmo tempo, uma visão geral dos dados e detalhes dos dados de interesse.

Distorção interativa - exibe porções de dados com um nível alto de detalhes (resolução), enquanto outros com um nível baixo. Em outras palavras, é como se a técnica exibisse os dados de interesse na perspectiva de uma lente de aumento, enquanto os outros dados são mostrados de maneira normal.

Ligação e seleção interativas - possibilitam a combinação de diferentes visualizações com o objetivo de tornar a exploração mais eficiente. Assim, é possível identificar correlações que dificilmente seriam percebidas utilizando apenas um tipo de visualização.

Considerando a classificação das técnicas de visualização de informação segundo Keim & Ward (2003), na metodologia proposta neste trabalho (a ser descrita no Capítulo 4) são utilizados três tipos de dados a serem visualizados, duas técnicas de visualização e duas técnicas interativas. Os dados visualizados podem ser bidimensionais (duas medidas para o gráfico X-Y), multidimensionais (três ou mais medidas para os gráficos de barra) e hierarquias (taxonomias). As técnicas de visualização utilizadas são o padrão 2D (Gráfico X-Y, Barras e Pizza) e hierárquica (visualização das taxonomias). Já as técnicas interativas utilizadas são a padrão (interação com a visualização, mudando-a dinamicamente de acordo com os objetivos) e a filtragem (filtragem interativa nos gráficos X-Y de acordo com valores de medidas ou itens de regras).

Conhecendo um pouco sobre o que é visualização de informação e a sua classificação, na próxima seção são apresentadas algumas formas de visualização de regras de associação.

3.3.2 *Visualização de Regras de Associação*

A visualização de informação pode ser utilizada para diversas finalidades, entre elas, para apresentar aos usuários os padrões obtidos da etapa de extração de padrões. Assim,

ao obter as regras de associação é possível utilizar diferentes formas de representação visual para apresentá-las ao usuário. Essas representações podem ser na forma de tabelas, matrizes multidimensionais, gráficos de barra, *Mosaic Plots*, *Double Decker Plots*, grades, grafos, coordenadas paralelas, *Enhanced Grid View*, *Tree View*, ou a combinação de diferentes técnicas (Keim & Ward, 2003).

Em Hofmann, Siebes, & Wilhelm (2000) são descritos *Mosaic Plots* e *Double Decker Plots*. A figura de um *Mosaic Plot* bidimensional, que ilustra um exemplo de visualização dos atributos X_1 e X_2 , pode ser observado no item (a) da Figura 3.2. Considerando uma regra de associação $X \Rightarrow Y$, na figura são ilustradas todas as combinações possíveis do antecedente X no *Mosaic Plot*, e o conseqüente Y (no caso da figura foi selecionado o valor y_2) no gráfico de barras. O tamanho dos retângulos no *Mosaic Plot* representa o suporte e o seu preenchimento representa a confiança. Por exemplo, se $X_1 = x_{11} \wedge X_2 = x_{22}$, pode-se observar que a confiança é alta e o suporte é baixo, já em $X_1 = x_{12} \wedge X_2 = x_{21}$ o suporte é o alto e a confiança é baixa.

O *Double Decker Plot*, que é uma variação do *Mosaic Plots*, pode ser observado no item (b) da Figura 3.2. A sua principal diferença é que os atributos do antecedente estão representados cada um em uma linha, na qual cada escala de cor representa um valor para este atributo. No caso da figura apresentada, X_1 pode assumir os valores x_{11} e x_{12} , representados pelo branco e preto, respectivamente. Já X_2 pode assumir os valores x_{21} , x_{22} , x_{23} e x_{24} , representados por uma escala de cor de tons de cinza (branco ao preto). No caso de dados binários, o branco representa “0” e o preto o “1”.

Figura 3.2: *Mosaic Plot* e a sua derivação *Double Decker Plots* (Hofmann, Siebes, & Wilhelm, 2000)

Em Ong, Ong, Ng, & Lim (2002) são descritas duas técnicas, *Enhanced Grid View* e *Tree View*, implementadas no protótipo CrystalClear. O *Enhanced Grid View* é representado por uma matriz, cujas células representam uma regra. As linhas da matriz indicam o suporte e as colunas a confiança, sendo que ambas estão ordenadas de forma crescente. Portanto, a última célula posicionada na direita e embaixo possui os maiores valores de suporte e confiança. Além disso, são utilizados cores e ícones para facilitar a visualização do usuário. Assim, para identificar as regras que se tornaram desinteressantes, por exemplo se o valor de suporte e confiança estão abaixo do desejado, são utilizados o ícone “-”. Para identificar as regras que se tornaram interessantes, por exemplo, se a adição de itens tornou os valores de suporte e confiança da regra acima dos valores desejados, é utilizado o símbolo “+”. Já para identificar as regras que foram geradas devido à adição de itens é utilizado o símbolo “#”. Também são utilizados os símbolos “↑” e “↓” para identificar as mudanças de suporte e confiança. As cores também podem ser usadas para facilitar a visualização das mudanças ocorridas nas regras. O usuário pode, por exemplo, designar uma cor para identificar as regras que se tornaram interessantes. Esta forma de representação pode ser vista na Figura 3.3 (a). Nesta figura é possível observar, que, ao posicionar o cursor sobre uma célula são exibidos os valores de suporte e confiança, além da regra.

Já a técnica *Tree View* exibe as regras de forma hierárquica, organizando as regras de acordo com o número de itens no antecedente e no conseqüente. As cores e ícones utilizados na técnica *Enhanced Grid View* também são utilizados para evitar a confusão na interpretação. Esta técnica pode ser visualizada na Figura 3.3 (b).

Blanchard, Guillet, & Briand (2003) representam subconjuntos de regras de associação em uma paisagem tridimensional. Essa paisagem é uma metáfora visual, na qual cada regra é representada por uma esfera sobre um cone, sendo que a esfera representa o suporte e o cone a confiança. Além disso, a posição da regra representa a implicação de intensidade, medida que avalia o quanto a regra é surpreendente, e a coloração proporciona uma idéia geral da qualidade da regra. Assim, uma regra que está representada por um cone e esfera grandes, na área da frente (“degraus” mais baixos), com uma coloração vermelha,

(a) *Enhanced Grid View*

(b) *Tree View*

Figura 3.3: *Enhanced Grid View* e *Tree View* implementadas no CrystalClear (Ong, Ong, Ng, & Lim, 2002)

representa uma regra surpreendente e com boa qualidade (confiança, suporte e intensidade altos). Já uma regra representada por um cone e esfera pequenos, na área do fundo (“degraus” mais altos), com uma coloração azul, representa uma regra não surpreendente e com baixa qualidade (confiança, suporte e intensidade baixos). A Figura 3.4 ilustra essa forma de representação.

Figura 3.4: Metáfora visual de Blanchard, Guillet, & Briand (2003)

Chakravarthy & Zhang (2003) descrevem uma abordagem para visualização em um espaço 3D, no qual cada linha representa um item do antecedente e cada coluna uma regra. Os nomes dos itens do antecedente são apresentados no lado direito e os nomes dos itens do conseqüente na parte inferior da coluna. Também são exibidos os valores de suporte e confiança, identificados por meio de cores e pelos símbolos “++” e “*”, respectivamente. Na Figura 3.5 é possível observar a visualização de 10 regras. Por exemplo, a primeira regra de

associação (mais à esquerda) exibida na figura é $Coat \ \& \ Lock \ \& \ Pump \implies Bike \ \& \ Eggs$ com suporte de 50% e confiança de 80%.

Figura 3.5: Abordagem utilizada por Chakravarthy & Zhang (2003)

Bruzzese & Buono (2004) utilizam a combinação das representações na forma de grafos e coordenadas paralelas. No item (a) da Figura 3.6 é possível observar a representação de aproximadamente 10000 regras de associação na forma de grafos. Já no item (b) as regras são apresentadas na forma de coordenadas paralelas.

Figura 3.6: Exemplos do modelo de Bruzzese & Buono (2004)

Em Techapichetvanich & Datta (2005) é descrita uma técnica híbrida, baseada nas representações na forma de matrizes e grafos, denominada VisAR. Esta técnica exibe os itens do antecedente e do conseqüente separadamente visando uma distinção mais clara dos seus itens. Os itens são representados por pontos e as regras por linhas verticais. Na Figura 3.7 pode ser observado um exemplo deste tipo de representação. Por exemplo, a primeira linha vertical da figura representa a regra de associação $cd \& rice \Rightarrow newspaper \& soya sauce$. Esta regra possui o maior suporte do conjunto de regras, uma vez que as regras estão ordenadas pelo suporte. Já as cores das linhas verticais representam a confiança. Também é possível ordenar as regras pela confiança. Neste caso, a cor das linhas verticais representa o suporte.

Figura 3.7: Técnica VisAR (Techapichetvanich & Datta, 2005)

Já em Yang (2005) são apresentadas as representações de *itemsets* frequentes e regras de associação representadas por coordenadas paralelas. Na Figura 3.8 (a) é exibida a representação dos *itemsets* frequentes *adbe*, *cdb* e *fg*. Já no item (b) pode ser observada a representação da regra de associação $ab \Rightarrow cd$. Observa-se que os itens do antecedente/-conseqüente são conectados por uma linha, enquanto que a relação entre eles é dada por uma seta.

Yang (2005) também apresenta a visualização de regras de associação utilizando taxonomias em coordenadas paralelas, exibida na Figura 3.9. As duas primeiras colunas representam o antecedente e a três últimas o conseqüente. A confiança é apresentada segundo uma escala de cor e o suporte pela espessura da linha. Já a taxonomia está representada pelos itens em cada coluna, sendo que os itens de uma taxonomia estão ligados por uma linha e estão organizados de acordo com a hierarquia (itens mais específicos estão

Figura 3.8: Exemplos de visualização de *itemsets* frequentes e regras de associação em coordenadas paralelas (Yang, 2005)

representados mais à direita).

Figura 3.9: Visualização com coordenadas paralelas Yang (2005)

Nesta seção foram apresentados os trabalhos relacionados com visualização de regras de associação encontrados durante a revisão bibliográfica realizada. Dentre estes trabalhos, apenas um trabalho, descrito em Yang (2005), visualiza regras de associação generalizadas. Entretanto, o trabalho de Yang (2005) utiliza apenas as tradicionais medidas suporte e confiança. As demais visualizações são de regras de associação simples e também utilizam apenas as medidas suporte e confiança. Assim, há a necessidade de desenvolver uma visualização para regras de associação generalizadas que disponibilize diferentes recursos, como diversas medidas e filtros, que realmente auxilie o usuário na exploração e avaliação de regras. A próxima seção apresenta trabalhos relacionados com avaliação de regras.

3.4 *Trabalhos Relacionados com Avaliação de Regras*

Nesta seção são apresentados alguns trabalhos relacionados com a avaliação de regras. Entre eles, encontram-se o aplicativo AIAS (Liu, Hsu, Chen, & Ma, 2000), o ambiente PEAR (Jorge, Poças, & Azevedo, 2002), o módulo de visualização de regras de associação descrito em Chakravarthy & Zhang (2003), o sistema VisAR (Techapichetvanich & Datta, 2005), o sistema I₂E (Yamamoto, Oliveira, Fujimoto, & Rezende, 2007; Yamamoto, Oliveira, & Rezende, 2008), e o ambiente RULEE e os seus módulos *ARInE* (Melanda, 2004), RULEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006) e RULEE-RAG (Fujimoto, Carvalho, & Rezende, 2007), desenvolvidos no LABIC-ICMC-USP.

3.4.1 *O aplicativo AIAS*

O aplicativo AIAS (*Association Interestingness Analysis System*), utiliza medidas subjetivas para avaliar regras de associação (Liu, Hsu, Chen, & Ma, 2000). O AIAS permite identificar regras interessantes, em especial regras inesperadas, a partir de um conjunto de regras de associação, aplicando o conhecimento do domínio para ordenar as regras segundo os tipos de inesperabilidade. Em seguida, o usuário pode utilizar os resultados das análises para visualizar as regras, salvar aquelas que são interessantes e remover as que não deseja.

O usuário deve fornecer o seu conhecimento do domínio, expresso por uma linguagem predefinida (descrita na Seção 3.2.2). O conhecimento pode ser expresso como impressão geral, conhecimento impreciso ou conhecimento preciso. A partir do conhecimento especificado, o sistema o utiliza para analisar as regras descobertas e calcular as medidas subjetivas: conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados. Calculados os valores das medidas, o usuário pode utilizá-los para visualizar as regras descobertas por diferentes aspectos (Liu, Hsu, Chen, & Ma, 2000), que estão descritas a seguir.

Tipos de regras potencialmente interessantes - as regras são categorizadas confor-

me o tipo de interesse, propiciando ao usuário a análise do conhecimento sob diferentes aspectos de interessabilidade.

Graus de interesse - as regras são separadas pelo grau de interesse (regras com maior grau de conformidade ou as mais inesperadas, por exemplo).

Itens interessantes - a apresentação dos itens interessantes pode ser considerada mais produtiva que a apresentação da regra como um todo.

Na Figura 3.10 é ilustrada uma tela do aplicativo *AIAS*, que utiliza medidas subjetivas. É possível observar um círculo na região central da tela, o qual indica o tipo de conhecimento prévio a ser considerado. Além disso, nos quadrantes ao redor do círculo são apresentadas as quatro medidas subjetivas implementadas, que são: conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados.

A coloração desses quadrantes indicam a existência de regras na faixa de valores especificados. Para cada quadrante colorido existe uma linha que o conecta ao círculo central, possuindo um número que indica a quantidade de regras presentes naquele quadrante para aquele tipo de conhecimento. Clicando no quadrante colorido são exibidas as regras que respeitam a faixa de valores especificada. Por exemplo, clicando no quadrante referente ao valor 0.5 da medida de conformidade (*Conform*), são exibidas as regras que respeitam a faixa de valores entre $0.5 \leq Conform < 0.75$. Além disso, em cada quadrante de faixa de valor é possível observar a utilização de *Listbox* para exibir os itens que estão ou não em conformidade com o tipo do conhecimento em análise.

3.4.2 O ambiente *PEAR*

O ambiente *PEAR* (*Post processing Environment for Association Rules*) tem como objetivo explorar grandes conjuntos de regras de associação (Jorge, Poças, & Azevedo, 2002). Inicialmente é apresentada uma página contendo um conjunto de regras (Figura 3.11), no qual o usuário pode navegar utilizando restrições e operadores. Algumas das restrições disponíveis são suporte e confiança mínimos. Com relação aos operadores, podem ser

Figura 3.10: Aplicativo AIAS (Liu, Hsu, Chen, & Ma, 2000)

citados operadores que selecionam regras com o mesmo antecedente ou conseqüente.

Figura 3.11: O ambiente PEAR (Jorge, Poças, & Azevedo, 2002)

O ambiente também permite a visualização das medidas suporte e confiança em gráficos X-Y e histograma (Figura 3.12). Os gráficos X-Y permitem observar as concentrações

de regras, segundo duas medidas, que neste caso são suporte e confiança. Já o histograma permite visualizar a distribuição das regras, de acordo com algum valor.

Figura 3.12: Visualização das medidas suporte e confiança em gráficos X-Y e histograma (Jorge, Poças, & Azevedo, 2002)

3.4.3 Módulo de visualização de regras de associação de Chakravarthy e Zhang

O módulo consiste na visualização das regras de associação no formato de tabelas e no espaço 3D, aplicando a técnica descrita na Seção 3.3.2. A visualização das regras de associação no formato de tabelas pode ser observado na Figura 3.13. No item (a) desta figura são exibidas todas as regras de associação e a opção de filtragem por meio da realização de consultas SQLs. Já no item (b) são exibidas as regras filtradas e a opção de ordenação de acordo com os valores de suporte e confiança.

Para a visualização das regras de associação, o primeiro passo é classificar as regras de associação de acordo com o número de itens no antecedente. No item (a) da Figura 3.14 é possível observar essa classificação, no qual o eixo X representa o número de itens no antecedente e o eixo Y o número de regras. Nesta figura também estão representados os valores de suporte e confiança.

Ao selecionar uma das colunas da Figura 3.14 (a) é gerada a visualização das regras de associação que possuem o número de itens determinados no antecedente. Esta visualização

HEAD	SYMBOL	BODY	CONFIDENCE	SUPPORT
Bike	=>	Lock	100%	50%
Lock	=>	Bike	67%	50%
Pump	=>	Lock	67%	50%
Lock	=>	Pump	67%	50%
Pump	=>	Coat	100%	75%
Coat	=>	Pump	100%	75%
Lock	=>	Coat	67%	50%
Coat	=>	Lock	67%	50%
Pump	=>	Lock, Coat	67%	50%
Lock	=>	Pump, Coat	67%	50%
Coat	=>	Pump, Lock	67%	50%
Pump, Lock	=>	Coat	100%	50%
Pump, Coat	=>	Lock	67%	50%
Lock, Coat	=>	Pump	100%	50%
Bike, Eggs	=>	Lock	100%	50%
Bike, Lock	=>	Eggs	50%	75%

(a) Filtragem das regras

HEAD	SYMBOL	BODY	CONFIDENCE	SUPPORT
Lock	=>	Bike	67%	50%
Lock	=>	Pump, Coat	67%	50%
Lock	=>	Coat	67%	50%
Lock	=>	Pump	67%	50%
Pump, Lock, Coat	=>	Bike, Eggs	80%	50%
Bike, Lock	=>	Eggs	80%	75%
Bike, Pump, Lock	=>	Eggs, Milk	80%	50%
Lock, Milk	=>	Coat	90%	75%
Bike, Lock, Eggs	=>	Coat, Milk	95%	85%
Pump, Lock	=>	Coat	100%	50%
Bike, Lock, Milk	=>	Coat	100%	50%
Lock, Coat	=>	Pump	100%	50%
Lock, Eggs, Coat	=>	Milk	100%	55%

(b) Ordenação das regras

Figura 3.13: Visualização de regras de associação no formato de tabelas (Chakravarthy & Zhang, 2003)

poede ser observada na Figura 3.14 (b). Como descrito na Seção 3.3.2, que apresenta essa abordagem de representação na Figura 3.5, cada linha representa um item do antecedente e cada coluna uma regra, sendo que os valores de suporte são representados por “++” e os valores de confiança por “**”.

(a) Número de itens no antecedente

(b) Visualização de regras de associação

Figura 3.14: Visualização 2D e 3D (Chakravarthy & Zhang, 2003)

3.4.4 O sistema VisAR

O sistema VisAR foi desenvolvido com o objetivo de reduzir a complexidade da visualização de um grande número de regras de associação (Techapichetvanich & Datta, 2005). Facilitando, desta forma, o entendimento do conjunto de regras. O sistema aplica

uma técnica, também chamada de VisAR, descrita na Seção 3.3.2, que foi desenvolvida baseada nas técnicas de matrizes e grafos.

Inicialmente, o usuário deve carregar as regras de associação no sistema, que realiza a contagem dos itens distintos nos antecedentes e conseqüentes, criando uma lista. Com a lista de itens criada, o usuário deve especificar os itens de interesse, selecionando-os e indicando se deseja utilizar a operação “OR” (ou) ou “AND” (e). Desta forma, o sistema pode filtrar as regras de acordo com os itens especificados. É importante ressaltar que no processo de filtragem, o sistema apenas verifica o antecedente da regra.

Em seguida, o usuário deve especificar se as regras deverão ser ordenadas pelos valores de suporte ou confiança. Assim, após realizadas estas tarefas, o usuário pode visualizar as regras de associação selecionadas. Na Figura 3.15 é possível observar um exemplo, no qual foram selecionados os itens *cd* e *rice*, utilizando a operação “OR” e a ordenação pelo suporte. Por exemplo, a primeira linha possui o mais alto suporte e representa a regra de associação *cd & rice & battery & soyasauce* $\implies$ *newspaper*.

Figura 3.15: Visualização de regras de associação no VisAR (Techapichetvanich & Datta, 2005)

Além disso, os usuários podem ver detalhes de cada regra de associação. Os valores de suporte e confiança são mostrados quando o usuário move o cursor sobre a linha vertical

que representa a regra.

3.4.5 O sistema I_2E e *RulEx*

O sistema I_2E e *RulEx*, apresentado em Yamamoto, Oliveira, Fujimoto, & Rezende (2007) e Yamamoto, Oliveira, & Rezende (2008), propõe inserir o usuário no processo de descoberta de regras de associação por meio do uso de técnicas de visualização de informação. O I_2E permite extrair interativamente *itemsets* freqüentes (a partir de uma implementação baseada no algoritmo Apriori) para gerar conjuntos compactos de regras de associação. O módulo possibilita também, entre os passos do algoritmo, a redefinição dos parâmetros do algoritmo, e a exploração visual dos *itemsets* freqüentes por meio de uma projeção 2D do espaço multidimensional dos mesmos (Figura 3.16).

Figura 3.16: O sistema I_2E (Yamamoto, Oliveira, & Rezende, 2008)

Na projeção apresentada, cada círculo representa um *itemset* freqüente, de modo que os mais similares são posicionados mais próximos entre si. Além disso, uma escala de cores mapeia os valores para uma característica do *itemset* freqüente (suporte ou agru-

pamento ao qual pertence, obtido com o algoritmo *k-means*). Para dar apoio a estas funcionalidades, são exibidos uma lista com os itens contidos pelos *itemsets* frequentes da visualização e um histograma que mapeia a distribuição de características dos mesmos (suporte ou identificador do agrupamento).

Os conjuntos de regras de associação gerados pelos *itemsets* selecionados no I₂E, podem ser explorados utilizando o sistema RulEx (Figura 3.17). Este módulo possibilita a exploração por meio da comparação visual de pares de regras, tomando como parâmetros oito medidas objetivas: Suporte, Confiança, Lift, Leverage, Certainty Factor, Jaccard, Gini Index e Yule's Q. A comparação é feita com o auxílio de uma visualização de coordenadas paralelas, na qual cada coordenada representa uma medida e uma poli-linha representa uma regra.

Figura 3.17: O sistema RulEx (Yamamoto, Oliveira, & Rezende, 2008)

3.4.6 O ambiente Rulee

Observando os ambientes relacionados com avaliação de regras de associação existentes, identificou-se a necessidade de um ambiente que forneça diferentes ferramentas ao usuário, facilitando a exploração das regras e a disponibilização de conhecimento. Com relação aos sistemas já apresentados nesta seção, é possível observar que três ambientes (PEAR (Jorge, Poças, & Azevedo, 2002), módulo de Chakravarthy & Zhang (2003) e VisAR (Techapichetvanich & Datta, 2005)) trabalham apenas com as tradicionais medidas suporte e confiança. O ambiente AIAS (Liu, Hsu, Chen, & Ma, 2000) utiliza apenas medidas subjetivas.

Assim, está sendo desenvolvido no LABIC o ambiente para exploração e disponibilização de regras RULEE (*Rule Exploration Environment*), que disponibiliza diferentes mecanismos para avaliação de regras. O RULEE auxilia a análise e disponibilização de regras de classificação, regressão e associação. Esse ambiente recebe como entrada arquivos com conjuntos de regras representadas na sintaxe padrão definida no DISCOVER (Prati, Baranauskas, & Monard, 2001; Pugliesi, Dosualdo, & Rezende, 2003; Melanda & Rezende, 2003), ambiente desenvolvido pelos pesquisadores do LABIC para apoiar todo o processo de mineração de dados. Por utilizar as regras na sintaxe padrão do DISCOVER e com utilização dos *scripts* de conversão dos diversos tipos de regras para a sintaxe padrão, também implementados no DISCOVER, podem ser inseridas no ambiente as saídas de uma grande variedade de algoritmos para extração de conhecimento de dados. Dentre esses algoritmos, encontram-se os mais utilizados na literatura, como o C4.5, CN2, *Cubist* e *Apriori*.

O ambiente RULEE foi baseado no protótipo desenvolvido por Paula (2003). Com a utilização desse protótipo, verificou-se a importância de um ambiente para exploração de regras, pois durante a etapa de pós-processamento e disponibilização do conhecimento é necessária a participação de usuários especialistas do domínio. Assim, o projeto foi reestruturado e está sendo implementado um ambiente para exploração e disponibilização

de regras, o qual possui o mesmo nome do protótipo, denominando-se RULEE ².

Uma vez que o conhecimento descoberto tenha sido armazenado no RULEE, podem ser então disponibilizados ao usuário não somente as regras, mas também métodos para avaliá-las. O RULEE apresenta um conjunto de medidas objetivas e subjetivas de avaliação que pode ser utilizado pelo usuário para ordenação e seleção de regras, podendo auxiliá-lo na identificação do conhecimento mais interessante.

A arquitetura definida para o RULEE, apresentada na Figura 3.18, é composta por: um conjunto de Repositórios (item (a)), uma Biblioteca de Acesso (item (b)), um Módulo de Gerenciamento (item (c)) e uma Interface (item (d)), que é composta pelos módulos de Exploração (item (e)) e Utilitários (item (g)).

Figura 3.18: Arquitetura geral do ambiente RULEE

O conjunto de Repositórios (item (a)) é constituído por: um Repositório de Dados Gerenciais, que armazena as informações sobre os usuários e os projetos; um Repositório de Conjuntos de Regras, que armazena os conjuntos de regras de associação, classificação e regressão nas respectivas sintaxe padrão; um Repositório de Medidas, que contém as informações sobre as medidas de avaliação de conhecimento, tanto objetivas quanto

²Projeto com apoio da FAPESP - Processo: 2006/04647-5.

subjetivas; um Repositório de Taxonomias, que armazena taxonomias de determinados domínios, utilizadas para generalizar regras de associação e para analisar conjuntos de regras generalizadas; e um Repositório de Dados, formado pelos conjuntos de dados utilizados na extração e teste de determinado conjunto de regras.

A Biblioteca de Acesso (item (b)) é formada por um conjunto de programas com funções específicas de acesso, manipulação e obtenção dos dados dos Repositórios, visando fazer a ligação entre o Módulo de Gerenciamento e os Repositórios da Base de Dados. Já o Módulo de Gerenciamento (item (c)) possui a função de gerenciar as informações sobre os usuários do RULEE e manter os Repositórios com valores consistentes.

A Interface (item (d)) é baseada na *Web*, com os objetivos de disponibilizar o conhecimento e permitir que um especialista do domínio explore o conjunto de regras de maneira fácil. Ela possui dois módulos auxiliares, o Módulo de Exploração (item (e)) e o Módulo de Utilitários (item (g)). O Módulo de Exploração é utilizado para explorar os conjuntos de regras de associação, classificação e regressão, exibidos no item (f). O Módulo de Utilitários é constituído por funções que auxiliam a interface a realizar diversas tarefas, como a comunicação com outros aplicativos, realizar o cálculo das medidas, entrada e saída (e/s) de informações, como pode ser observado no item (h).

Para o ambiente prover um fácil acesso às regras e aos valores dessas medidas e, também, permitir que as medidas sejam utilizadas de forma combinada na análise das regras, são disponibilizados filtros para ordenação dos conjuntos de regras. Além disso, o ambiente RULEE está sendo implementado para permitir, com pequenas alterações em seu código fonte, a incorporação de novas medidas de avaliação pelos pesquisadores e novos módulos de pós-processamento. Assim, com a utilização desse ambiente os pesquisadores podem realizar experimentos com novas medidas de avaliação sem despendar muito tempo com implementações. Além disso, uma característica importante do ambiente RULEE é a facilidade de acesso ao conhecimento descoberto, realizado por meio de uma interface para disponibilização do conhecimento baseada na *World Wide Web* (WWW). Dessa maneira, algumas características importantes da WWW são incorporadas ao ambiente, como a independência de plataforma e a facilidade de utilização pelos usuários independente de

sua localização física e da ferramenta de extração utilizada.

Também estão disponíveis nesse ambiente, como Módulo de Exploração de regras de associação segundo a Figura 3.18, os módulos: RULEE-SACT (Martins, 2006) para construção semi-automática de taxonomias, *ARInE* (Melanda, 2004) para análise com medidas objetivas, RULEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006) para análise com medidas subjetivas e RULEE-RAG (Fujimoto, Carvalho, & Rezende, 2007) para auxiliar a avaliação de regras de associação generalizadas. Estão detalhados, a seguir, os módulos *ARInE*, RULEE-SEAR e RULEE-RAG, por estarem mais relacionados com este trabalho.

O módulo ARInE

O *ARInE* (*Environment for Association Rules Interactive Exploration*) é um módulo do ambiente RULEE que foi projetado e implementado para disponibilizar um ambiente *Web* com o objetivo de apoiar a avaliação de regras de associação (Melanda, 2004). Essa avaliação é baseada na metodologia proposta por Melanda (2004), que realiza a análise de regras de associação fundamentadas em análises gráficas de medidas objetivas. Com apoio dessa metodologia é possível filtrar as regras de associação, identificando pequenos grupos de regras para serem avaliados por um especialista.

A primeira etapa desse processo de identificação de regras potencialmente interessantes possui como entrada um conjunto de regras de associação resultante da etapa de extração de padrões. O usuário pode definir o seu conjunto de regras foco utilizando métodos de consulta no conjunto de regras. Nessas consultas podem ser especificados os itens que o usuário tem interesse e pode ser definido se esses itens devem ocorrer no antecedente, no conseqüente ou em qualquer lado da regra. O resultado da consulta é apresentado ao usuário, que pode optar por realizar uma nova consulta, salvar ou excluir regras.

Definido o conjunto de regras a ser analisado, a próxima etapa consiste na aplicação de filtros utilizando medidas objetivas, descritas em (Melanda, 2004). A aplicação de filtros representa o conceito central da metodologia do *ARInE*, que é identificar pequenos

conjuntos de regras de maneira que o usuário possa interpretá-los e identificar as regras que lhe são interessantes. Estes pequenos conjuntos podem ser, por exemplo, identificados visualmente observando-se as áreas em destaque em um gráfico de dispersão. Contudo, nem sempre a identificação destes conjuntos é trivial, como no caso no qual os pontos estão dispersos. Nesse caso, é necessário o uso de outros métodos de análise. Para tanto, são disponibilizadas outras duas maneiras de análise também baseadas no estudo de gráficos: a distribuição de frequência e a análise de Pareto.

Na análise da distribuição de frequência, os valores de uma dada medida são discretizados, sendo verificados o número de regras para cada intervalo. Já a análise de Pareto, que é baseada no Princípio de Pareto, afirma que: “para a maioria dos fenômenos naturais, 80% das conseqüências são originadas por apenas 20% das causas”. No âmbito da avaliação de regras de associação, o Princípio de Pareto pode ser interpretado da seguinte maneira: “apenas 20% das regras são responsáveis por 80% do valor acumulado de uma medida”. Desse modo, é possível priorizar a análise de regras considerando-se apenas as regras que mais contribuem para uma dada medida. Na análise de Pareto, é gerado o gráfico da soma cumulativa dos valores da medida para cada regra considerada. Então o gráfico é apresentado ao usuário que o interpreta e define o percentual limite de corte para seleção das regras. Como referência, é destacado o percentual de 80% no gráfico.

O módulo RuLEE-SEAR

O módulo RULEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006) foi desenvolvido para possibilitar a exploração de regras de associação com medidas subjetivas. Sinoara (2006) propôs uma metodologia para identificação de regras de associação interessantes que combina o uso de medidas objetivas e subjetivas. Essa metodologia é ilustrada na Figura 3.19, sendo dividida em quatro etapas: análise objetiva, captação do conhecimento e dos interesses do especialista, processamento da avaliação e análise com medidas subjetivas. A primeira etapa é realizada no módulo *ARInE* (Melanda, 2004) e as etapas seguintes (etapas 2, 3 e 4) são realizadas no RULEE-SEAR.

Figura 3.19: Metodologia para identificação de regras de associação interessantes (Sinoara, 2006)

Na etapa de análise objetiva, executada no *ARInE*, são realizadas consultas para selecionar um conjunto de regras foco. Além disso, também são realizadas análises com as medidas objetivas para filtrar as regras e obter um subconjunto com regras potencialmente interessantes.

No RULEE-SEAR é realizada a etapa de captação do conhecimento e dos interesses do especialista do domínio, na qual ele avalia cada regra desse subconjunto. Paralelamente a essa avaliação é realizada a eliminação de conhecimento irrelevante, como parte do processamento da avaliação. Quando o especialista classifica uma regra como conhecimento irrelevante, essa regra e as regras similares a ela são eliminadas do conjunto foco. Finalizada a avaliação inicia-se a segunda fase do processamento, o cálculo das medidas subjetivas. Essas medidas são calculadas para as regras do conjunto foco que não foram eliminadas com base nas regras avaliadas (excluindo aquelas classificadas como irrelevantes). Com os valores das medidas subjetivas pode-se iniciar a análise com essas medidas, última etapa da metodologia. Durante essa análise o especialista utiliza as me-

didadas subjetivas para guiá-lo na exploração das regras resultantes no conjunto foco e na identificação das regras interessantes.

O módulo RuLEE-RAG

O módulo de regras de associação generalizadas, denominado RuLEE-RAG (Fujimoto, Carvalho, & Rezende, 2007), foi desenvolvido para viabilizar a aplicação da metodologia proposta por Carvalho (2007), na qual utiliza taxonomias para generalizar conjuntos de regras de associação. O RuLEE-RAG permite a inserção e exploração dos conjuntos de regras de associação generalizadas na etapa de pós-processamento do conhecimento. O objetivo do módulo RuLEE-RAG é auxiliar a avaliação das regras de associação generalizadas por meio da exploração interativa.

O módulo possui como entrada o identificador do conjunto de regras de origem e os arquivos de regras de associação generalizados, de taxonomias de itens específicos, de dados e de taxonomias. Em seguida, os dados são salvos na base de dados do ambiente RuLEE. O módulo permite a exploração das regras de associação generalizadas inseridas no ambiente, exibindo para cada projeto os conjuntos de regras de associação generalizadas existentes. Quando o usuário seleciona um determinado conjunto de regras, é possível escolher entre exibir as regras ou realizar consultas SQL. O módulo exibe as regras específicas e generalizadas, bem como a contribuição dos itens específicos e as regras de origem da regra generalizada.

É importante ressaltar que o módulo RuLEE-RAG permite apenas a visualização da regra na forma textual e a exploração com algumas medidas objetivas (Suporte, Confiança, Lift, Convicção e Leverage). Com a utilização do módulo RuLEE-RAG, foi identificada a necessidade de um ambiente que forneça diferentes mecanismos que realmente auxilie o usuário na exploração das regras de associação generalizadas, de modo a melhorar a compreensão e facilitar a identificação de conhecimento interessante. Assim, observando essa necessidade, este trabalho propõe uma metodologia, a ser descrita no Capítulo 4, que utiliza medidas de avaliação objetivas e subjetivas em conjunto com técnicas de visualização

de informação. Para viabilizar a aplicação desta metodologia, também foi desenvolvido o módulo denominado RULEE-GARVIS, que passa a integrar o RULEE.

3.5 *Considerações Finais*

Neste capítulo foram discutidos algumas abordagens utilizadas para avaliação de regras, como medidas de avaliação de conhecimento e técnicas de visualização de informação. Também foram apresentados alguns trabalhos relacionados com a avaliação de regras. Analisando esses trabalhos, percebe-se a necessidade de auxiliar o usuário não apenas na identificação do conhecimento interessante, mas também na compreensão das regras. Os trabalhos de Jorge, Poças, & Azevedo (2002), Chakravarthy & Zhang (2003) e Techapichetvanich & Datta (2005), utilizam apenas as tradicionais medidas suporte e confiança. O trabalho de Yamamoto, Oliveira, Fujimoto, & Rezende (2007); Yamamoto, Oliveira, & Rezende (2008) utiliza outras medidas além de suporte e confiança, entretanto o foco do trabalho é a etapa de extração de padrões. Já o trabalho de (Liu, Hsu, Chen, & Ma, 2000) utiliza apenas medidas subjetivas.

O ambiente RULEE, em conjunto com os seus módulos *ARInE*, RULEE-SEAR e RULEE-RAG, disponibiliza diferentes mecanismos para avaliação de regras. Apesar desses diferentes mecanismos implementados, foi identificada a necessidade de um módulo que auxilie a compreensão e a identificação de regras de associação generalizadas interessantes. Isso porque o RULEE-RAG, que era único módulo até então usado para auxiliar a exploração de regras de associação generalizadas, possui uma variedade de medidas disponíveis limitadas e apresenta as regras apenas na sua forma textual. Como a apresentação de muitas regras na sua forma textual pode dificultar a compreensão e a identificação de conhecimentos interessantes, este trabalho foi proposto com o intuito de suprir essa necessidade. Assim, foi proposta uma metodologia que utiliza técnicas de visualização de informação em conjunto com técnicas de avaliação por consulta e medidas de avaliação do conhecimento. Além disso, foi desenvolvido o módulo RULEE-GARVIS, que viabiliza a aplicação da metodologia.

Metodologia para Incrementar a Compreensibilidade de Regras de Associação Generalizadas e Apoiar a Identificação de Conhecimento Interessante

4.1 *Considerações Iniciais*

Como mencionado nos capítulos anteriores, a grande quantidade de padrões gerados por alguns algoritmos de extração de regras de associação dificulta a compreensão e a identificação de conhecimento interessante. Para tal, existem diferentes abordagens que auxiliam a avaliação de regras de associação, como avaliação por consulta, aplicação de medidas de avaliação do conhecimento, generalização, visualização. As medidas de avaliação de conhecimento podem ser objetivas, dependendo apenas da estrutura dos padrões e dos

dados utilizados, ou subjetivas, considerando o conhecimento e o interesse dos usuários. A generalização utiliza o conhecimento do domínio, que pode ser representado hierarquicamente na forma de taxonomias, para permitir a descoberta de um conhecimento mais geral. Já as técnicas de visualização procuram melhorar a compreensão, aproveitando a capacidade humana de interpretação visual.

Com base na revisão bibliográfica, neste capítulo é apresentada a metodologia proposta com o objetivo de auxiliar a compreensão e a identificação de conhecimento interessante em regras de associação generalizadas. Para isso são utilizadas medidas de avaliação objetivas e subjetivas em conjunto com técnicas de visualização. A utilização de medidas objetivas e subjetivas em conjunto visam maximizar as suas vantagens e minimizar as suas desvantagens, visto que apenas o uso de medidas objetivas pode não ser suficiente na identificação de regras interessantes, uma vez que os objetivos e o conhecimento dos usuários não são considerados. A utilização exclusiva de medidas subjetivas também é complicada, pois o usuário não possui uma medida que possa servir de filtro para a seleção do conjunto de regras a ser avaliado.

Para viabilizar a utilização da metodologia proposta, foi desenvolvido um módulo denominado RULEE-GARVIS. Este módulo é integrante do ambiente RULEE, descrito na Seção 3.4.6. É importante destacar que este trabalho adapta o fluxo de análise proposto por Sinoara (2006), que propôs uma metodologia que combina a utilização de medidas objetivas e subjetivas para identificação de regras de associação interessantes. Em Sinoara (2006), a análise com medidas objetivas é realizada no *ARInE* e com medidas subjetivas no RULEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006). O módulo *ARInE* e RULEE-SEAR não foram utilizados, pois não consideram os aspectos particulares de regras de associação generalizadas. Além disso, o *ARInE* não possibilita a interação direta do usuário com os gráficos e o RULEE-SEAR não apresenta nenhum gráfico disponível. Assim, este trabalho foi desenvolvido aproveitando algumas funcionalidades disponíveis no RULEE, como o cálculo de medidas e inserção de regras.

Um dos principais diferenciais deste trabalho é a utilização de técnicas de visualização para melhorar a compreensibilidade. A visualização proposta é uma visualização

interativa, como no caso do gráfico X-Y, no qual o usuário pode “clique” na regra, filtrar dinamicamente por valores de medidas ou estrutura da regra (itens do antecedente ou conseqüente) ou gerar diferentes gráficos combinando diferentes medidas, além de poder avaliar uma regra considerando a hierarquia de uma taxonomia, disponibilizada por meio de uma visualização que faz uma analogia com a estrutura de diretórios. Também está disponível a análise de regras reduntantes, complementos e exceções, para que o usuário tenha disponível diferentes mecanismos que auxiliem a compreensão e identificação de conhecimentos interessantes.

4.2 *Metodologia Proposta*

A metodologia proposta permite que o usuário analise objetivamente o conjunto de regras de associação generalizados por meio de consultas SQLs e análises gráficas, selecionando um subconjunto a ser avaliado pelo especialista do domínio. Com a avaliação das regras selecionadas, medidas subjetivas podem ser calculadas e a exploração com as medidas subjetivas iniciada. Dessa forma, o usuário pode analisar o conjunto de regras observando diferentes aspectos. A metodologia, ilustrada na Figura 4.1, está dividida em quatro etapas, descritas a seguir.

Análise objetiva (1) - a metodologia considera que o usuário já possui as taxonomias e o conjunto de regras de associação generalizadas. Inicialmente são realizadas consultas SQL selecionando determinados atributos e/ou valores de medidas objetivas¹ para selecionar um conjunto foco. Caso o usuário não tenha interesse em um item em específico, pode ser selecionado todo o conjunto de regras como conjunto foco. Com apoio de análises gráficas nas regras do conjunto foco utilizando algumas medidas objetivas², são filtradas o conjunto de regras de associação identificando subconjunto de regras potencialmente interessantes (RPI). As análises gráficas desta etapa são realizadas de forma

¹Disponíveis medidas objetivas, que possuem intervalos abertos e fechados, descritas em Melanda (2004) e algumas medidas descritas em Carvalho et al. (2007).

²As medidas utilizadas nas análises gráficas são: IS/Cosine, Jaccard, ϕ -coefficient, Piatetsky-Shapiro's, Kappa, Laplace, Gini Index, Added Value, Certainty Factor, Confidence e Support. Estas medidas foram selecionadas por possuírem um intervalo fechado e estão descritas no trabalho de Carvalho et al. (2007), que estudou as medidas mais apropriadas para avaliação de regras de associação generalizadas.

no caso da regra avaliada ser generalizada, além das regras redundantes, complementos e exceções. No caso das regras redundantes, dos complementos e das exceções, o usuário pode adicionar a(s) regra(s) desejada(s) no subconjunto RPI e avaliá-la(s). Para um melhor entendimento dos termos complemento, redundância e exceção, a seguir estão suas definições. A definição de complemento é proposta neste trabalho pela autora e são utilizadas as definições de Zaki (2004) e Gonçalves et al. (2005) para redundância e exceção, respectivamente.

Um complemento pode ser definido como: dado um item generalizado em uma regra e uma taxonomia, os complementos dessa regra são todas as regras que possuem a mesma estrutura, diferenciando apenas no item generalizado. Os itens generalizados desses complementos possuirão o mesmo ancestral na taxonomia em relação ao item generalizado dado como referencial. Assim, seja $R = \{r_1, \dots, r_l\}$ um conjunto de regras e $X = \{x_1, \dots, x_k\}$ um conjunto de taxonomias. Dado que $r_i, r_j \in R$, r_i é complemento de r_j se $((r_i.LHS = r_j.LHS \wedge r_i.RHS \setminus r_j.RHS = w \wedge r_j.RHS \setminus r_i.RHS = w') \vee (r_i.RHS = r_j.RHS \wedge r_i.LHS \setminus r_j.LHS = w \wedge r_j.LHS \setminus r_i.LHS = w'))$ e w, w' possuem o mesmo ancestral na taxonomia x_h . O complemento é simétrico, então r_i é complemento de r_j e r_j é complemento de r_i . Exemplo: considerando uma taxonomia que possua *doces* como ancestral e *chocolate*, *sobremesa_cremosa_chocolate* e *biscoito* como filhos, dada a regra referencial *chocolate* $\Rightarrow$ *leite_moca*, são complementos dessa regra *sobremesa_cremosa_chocolate* $\Rightarrow$ *leite_moca* e *biscoito* $\Rightarrow$ *leite_moca*.

Uma regra redundante, segundo Zaki (2004), pode ser definida como: seja $S = \{s_1, \dots, s_o\}$ um conjunto de regras, que possuem todos suportes e confiâncias iguais. Dado $s_i, s_j \in S$, podemos dizer que s_j é redundante se existe algum s_i , tal que s_i é mais geral que s_j ($s_i \leq s_j$). Lembrando que Zaki (2004) considera regras com menos itens mais gerais. Exemplo: dada a regra $(r)chocolate \Rightarrow leite_moca$, uma possível regra redundante candidata

é $(rrc)chocolate \ \& \ pao \Rightarrow \ leite_moca$. A regra (rrc) será redundante se $\text{suporte}(rrc) = \text{suporte}(r)$ e $\text{confiança}(rrc) = \text{confiança}(r)$. A condição de igualdade considerada é a comparação de três casas decimais (por exemplo, os valores 0.2345 e 0.2349 são considerados iguais por possuírem as três primeiras casas decimais semelhantes).

Uma *exceção*, segundo Gonçalves et al. (2005), pode ser definida como: seja $A \Rightarrow B$ uma regra e $Z = \{z_1, \dots, z_p\}$ um conjunto de itens, tal que $Z \cap A \cap B = \emptyset$. Uma regra candidata a exceção possui a forma $A \ \& \ Z \Rightarrow B$. No caso da presença de Z enfraquecer a regra em comparação com a regra $A \Rightarrow B$, $A \ \& \ Z \Rightarrow B$ é considerada uma exceção negativa. Caso contrário, uma exceção positiva. Exemplo: dada a regra $(r)chocolate \Rightarrow \ leite_moca$, uma possível regra de exceção candidata é $(rec) \ chocolate \ \& \ pao \Rightarrow \ leite_moca$. A regra de exceção candidata será uma exceção positiva se $\text{suporte}(rec) < \text{suporte}(r)$ e $\text{confiança}(rec) > \text{confiança}(r)$. Caso $\text{confiança}(rec) < \text{confiança}(r)$ a regra será uma exceção negativa.

Assim, nesta etapa podem ser utilizados diferentes mecanismos para auxiliar o usuário na avaliação das regras. Estão disponíveis a visualização das regras em gráficos X-Y ou de barras, a visualização da taxonomia em um formato de árvore que faz analogia com a estrutura de um diretório, a visualização da contribuição dos itens específicos, além da possibilidade de análise das regras de origem, complemento, exceção e redundância. Finalizadas as avaliações das regras do subconjunto RPI, pode ser iniciada a próxima etapa da metodologia.

Processamento subjetivo (3) - nesta etapa, para as regras do conjunto foco que não foram eliminadas, são calculadas as medidas subjetivas conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados, definidos por Liu et al. (2000). Desta forma, é possível realizar uma análise com as medidas subjetivas para auxiliar a identificação de possíveis regras interessantes não encontradas anteriormente.

Análise com medidas subjetivas (4) - durante a exploração das regras do conjunto foco resultante utilizando as medidas subjetivas calculadas na Etapa 3, o usuário possui o auxílio de análises gráficas e da taxonomia. Os gráficos nesta etapa utilizam as medidas subjetivas e estão disponíveis gráficos X-Y e de barras. Com essas visualizações pretende-se aumentar a compreensibilidade das regras e facilitar a identificação do conhecimento interessante, pois fornece uma opção visual e interativa de exploração além da forma textual que pode ser cansativa. É importante destacar que a navegação no conjunto foco resultante é realizada de acordo com os objetivos do usuário no momento da avaliação.

Um ponto relevante a ser ressaltado é que neste trabalho é adaptado o fluxo de análise com medidas objetivas e subjetivas proposto por Sinoara (2006). Assim, o diferencial deste trabalho está na integração de técnicas de visualização de informação para aumentar a compreensibilidade das regras, além de incorporar aspectos particulares de regras de associação generalizadas e a possibilidade de análise de regras redundantes, de exceção e complementos. Na seção seguinte são realizadas algumas considerações sobre a metodologia.

4.3 Considerações sobre a Metodologia Proposta

Na metodologia proposta, considera-se que o conjunto de regras de associação foi generalizado utilizando a metodologia de Carvalho (2007). Como mencionado na Seção 2.3.2, Carvalho, Rezende, & Castro (2007) realizam a generalização de regras de associação na etapa de pós-processamento do conhecimento utilizando o algoritmo $APRA_{alg}$, a partir de um conjunto de regras de associação e uma taxonomia. Para a aplicação da metodologia de Carvalho (2007), também foi desenvolvido o módulo RULEE-RAG, descrito na Seção 3.4.6, que permite a inserção do conjunto de regras de associação generalizados e a sua exploração textual.

Assim, o RULEE-GARVIS utiliza a funcionalidade de inserção do módulo RULEE-RAG (Fujimoto, Carvalho, & Rezende, 2007), que possui como entrada os arquivos de regras de associação generalizadas, de taxonomias de itens específicos, de dados e de

taxonomias. Além disso, é necessário informar o identificador do conjunto de regras de associação de origem. Algumas outras considerações estão listadas a seguir:

- a metodologia permite a análise de regras generalizadas e específicas;
- as medidas objetivas utilizadas na consulta SQL da Etapa 1 da metodologia, que visa a seleção do conjunto foco, são as medidas cadastradas no ambiente RULEE, que são as apresentadas na Seção 3.2 e as descritas em (Melanda, 2004);
- as medidas objetivas utilizadas nas análises gráficas são as medidas apresentadas na Seção 3.2.1. Estas medidas são algumas das medidas utilizadas por Carvalho, Rezende, & Castro (2007), sendo selecionadas por estarem cadastradas no RULEE e por possuírem um intervalo fechado (bem definido). É importante ressaltar que, neste projeto, foram utilizadas medidas com intervalo fechado para as visualizações da metodologia, pelo fato de serem mais facilmente mapeadas para uma representação visual 2D;
- as medidas subjetivas utilizadas na metodologia proposta são conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados, definidos por Liu, Hsu, Chen, & Ma (2000);
- as análises gráficas podem ser realizadas em Gráficos X-Y, que permitem a interação do usuário com os gráficos. Além desse gráfico, também estão disponíveis gráficos de barra e pizza implementados utilizando a biblioteca JFreeChart³;
- o usuário tem a opção de encerrar a exploração no momento que achar necessário. Com a exploração encerrada, o usuário pode visualizar textualmente e graficamente as regras interessantes e/ou a avaliação das regras (no caso do usuário ter encerrado após a realização da Etapa 2, de avaliação das regras do subconjunto RPI).

Como citado anteriormente, este trabalho adapta o fluxo de análise proposto por Sinoara (2006). Em Sinoara (2006) a análise objetiva é realizada no *ARInE* (Melanda,

³Biblioteca *open source* de Gráficos implementada em JAVA. Disponível em: <http://www.jfree.org/jfreechart/>

2004) e a subjetiva no RULEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006), lembrando que o fluxo de análise foi proposto para regras de associação, não considerando os aspectos particulares de uma regra de associação generalizada. Neste trabalho, tanto a análise objetiva como a subjetiva são realizadas em um único módulo denominado RULEE-GARVIS. O módulo *ARInE* não foi utilizado, pois não considera os aspectos particulares de uma regra de associação generalizada. Além disso, no *ARInE* todos os gráficos são estáticos, não havendo a possibilidade de interação direta com nenhum deles. Em relação ao RULEE-SEAR, alguns métodos puderam ser aproveitados no RULEE-GARVIS, como o cálculo das medidas subjetivas.

4.4 *Módulo de Visualização de Regras de Associação Generalizadas*

- RuLEE-GARVis

O módulo RULEE-GARVIS foi desenvolvido com o intuito de validar a metodologia proposta. O módulo RULEE-GARVIS considera que as regras já foram generalizadas e que foram inseridas pelo módulo RULEE-RAG (Fujimoto, Carvalho, & Rezende, 2007). Assim, para a sua utilização são necessários o cumprimento de alguns requisitos discutidos a seguir.

4.4.1 *Requisitos para a utilização do Módulo RuLEE-GARVis*

Para a perfeita utilização das funcionalidades do módulo RULEE-GARVIS são necessários alguns requisitos, como:

- possuir conexão com a *Internet* (preferencialmente banda larga) e um *browser* que suporte o *plugin* JAVA;
- possuir um usuário (login) no ambiente RULEE ⁴;

⁴Disponível em: <http://143.107.232.107/rulee/index.html>.

- generalizar as regras e inserir as taxonomias e as regras generalizadas no RULEE pelo módulo RULEE-RAG (Fujimoto, Carvalho, & Rezende, 2007);
- possuir no mínimo a versão 1.6 do JAVA no computador;

Atendendo os requisitos listados acima, o usuário estará apto a utilizar o módulo RULEE-GARVIS. Na próxima seção são apresentadas as funcionalidades do módulo.

4.4.2 Funcionalidades do Módulo RULEE-GARVIS

O módulo RULEE-GARVIS foi desenvolvido aproveitando as vantagens oferecidas pelo ambiente RULEE e pelos seus módulos RULEE-RAG (Fujimoto, Carvalho, & Rezende, 2007) e RULEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006). Assim, foram aproveitadas algumas tabelas da base de dados e métodos de classes já existentes. Para o seu funcionamento foram acrescentadas uma tabela na base de dados, uma classe e alguns métodos em classes já existentes. Além disso, foram implementadas 22 páginas com extensão *html* para exibir a interface. Na Figura 4.2 é apresentado o diagrama de caso de uso do módulo RULEE-GARVIS para uma visão geral das funcionalidades.

A seguir estão descritas as funcionalidades do módulo RULEE-GARVIS. A descrição completa da implementação está em Fujimoto & Rezende (2008a).

- **Opções ao iniciar o módulo:** criar uma nova exploração, continuar uma exploração já existente ou visualizar uma exploração já encerrada.
- **Criar Exploração:** o usuário deve selecionar um projeto e um conjunto de regras. Ao escolher o conjunto de regras, algumas informações como o número de regras e o lado da generalização são exibidos. Lembrando que é possível realizar generalização no lado esquerdo (antecedente), direito (conseqüente) ou em ambos os lados. A generalização do lado esquerdo representa o relacionamento entre categorias/classes de itens e itens específicos, do lado direito representa o relacionamento entre itens específicos e categorias/classes de itens e a generalização de ambos os lados repre-

Figura 4.2: Diagrama de Caso de Uso do RULEE-GARVis

senta o relacionamento entre categorias/classes de itens (Carvalho, 2007). Definido o conjunto, o usuário deve selecionar as regras que formarão o conjunto foco.

- **Continuar Exploração:** o usuário pode continuar um processo de exploração iniciado. Ao selecionar o processo, o sistema exibe informações como código do projeto, conjunto de regras, identificador da sessão de análise, número de regras do conjunto, usuário responsável da exploração e as etapas realizadas para auxiliar a identificação do processo desejado.
- **Selecionar Conjunto Foco:** o usuário pode selecionar todo conjunto de regras ou realizar consulta. No caso de realizar consulta, o usuário pode filtrar as regras por algum item, valores de medidas objetivas cadastradas no ambiente RULEE. Após a consulta ser realizada, o sistema possibilita a edição da consulta, a visualização do resultado ou a visualização do resultado em relação a todo conjunto de regras.

- **Selecionar Regras Potencialmente Interessantes:** a seleção acontece por meio da utilização da visualização de medidas objetivas (apresentadas na Seção 3.2) no gráfico X-Y, com auxílio de filtros e a visualização da taxonomia. As regras que constituirão o conjunto de regras potencialmente interessantes (RPI) são as regras selecionadas nesta fase. Ao final, o usuário possui a opção de iniciar a avaliação das regras ou encerrar a exploração.
- **Avaliar regras do subconjunto RPI:** na avaliação das regras do subconjunto RPI são apresentadas todas as regras do subconjunto, uma por uma. As regras devem ser avaliadas como inesperadas, úteis, óbvias do domínio, de conhecimento prévio do especialista ou irrelevantes. Caso o usuário avalie uma regra como irrelevante, ele terá que indicar qual(is) item(s) que tornaram essa regra irrelevante. Durante a avaliação, o usuário pode escrever um comentário sobre a avaliação dessa regra e deve indicar se a regra é interessante ou não. Como informações adicionais, no caso da regra ser generalizada, o sistema exibe as regras de origem e a contribuição dos itens específicos. Paralelamente à exibição textual das regras, é possível visualizar a taxonomia e gráficos de barra e X-Y. Além disso, são exibidas as regras de complemento, exceção e redundância, no caso destas existirem. Se o usuário possuir interesse por alguma dessas regras, ele pode adicionar essa regra ao subconjunto RPI para avaliá-la. Após finalizada a avaliação das regras, é possível calcular as medidas subjetivas ou encerrar a exploração. Caso o usuário deseje calcular as medidas subjetivas, as seguintes medidas são calculadas: conformidade ($confm_{ij}$), antecedente inesperado ($unexpCond_{ij}$), consequente inesperado ($unexpConseq_{ij}$), antecedente e consequente inesperados ($bsUnexp_{ij}$), definidas em Liu, Hsu, Chen, & Ma (2000).
- **Explorar com medidas subjetivas:** em um primeiro momento, as regras exibidas textualmente são divididas de acordo com a sua avaliação, inesperadas/úteis ou óbvias do domínio/conhecimento prévio. Paralelamente são exibidas a taxonomia e o gráfico X-Y das regras com as medidas objetivas e coloridas de acordo com a sua avaliação. Ao selecionar uma regra, o usuário pode escolher uma ou mais das medidas subjetivas e definir os seus intervalos. Definidas as medidas subjetivas e os

intervalos, são exibidas as regras que satisfazem os intervalos das medidas subjetivas selecionadas. Nesta fase estão disponíveis ao usuário a visualização da taxonomia e dos gráfico de barras e X-Y com as medidas subjetivas. No gráfico de barras, o usuário pode visualizar utilizando apenas para as medidas selecionadas ou todas as quatro medidas subjetivas. Assim, o usuário pode identificar regras interessantes não encontradas previamente e adicionar comentários sobre as regras.

- **Visualizar Regras Interessantes:** o usuário pode visualizar as regras na sua forma textual ou utilizando gráficos de barra e pizza. Esses gráficos permitem a análise do número de regras interessantes em relação ao número de regras úteis, inesperadas, prévias, óbvias, generalizadas, específicas ou o total de regras.
- **Visualizar as Explorações Encerradas:** o usuário possui a opção de analisar a avaliação (no caso de ter sido realizada) e/ou visualizar as regras interessantes. Na visualização da avaliação, o sistema exibe as regras textualmente e possibilita uma visão geral das regras avaliadas por meio de um gráfico de barras. Esse gráfico exibe o número de regras avaliadas como inesperadas, úteis, prévias, óbvias, irrelevantes, interessantes e o total de regras avaliadas. Já a visualização de regras interessantes é parecido com o descrito no item anterior.

A fim de viabilizar a implementação do gráfico X-Y, a seguir são apresentados dois dos algoritmos que compõem o *GARVisalg*, que viabiliza a visualização de regras de associação generalizadas no espaço 2D. No Algoritmo 1 é calculado a posição inicial das regras. Inicialmente são verificadas quais as medidas selecionadas (são selecionadas como padrão o suporte e a confiança) para os eixos x e y, respectivamente. Para calcular a posição da regra no eixo x, o algoritmo faz uma chamada ao *CalculaPosicaoRegraEixoX($r, MedidaX$)*, sendo que r é a regra e *MedidaX* é a medida selecionada para o eixo x. Neste algoritmo, são obtidos os valores mínimo e máximo para a medida e para a área disponível de plotagem. Com o valor da medida para a regra, é realizado o cálculo para determinar a posição no eixo x para a plotagem da regra. De maneira análoga, é realizado o cálculo para o eixo y. Após o cálculo da posição da regra na área de plotagem, é verificada

a necessidade de *jitter*, uma perturbação inserida para evitar a sobreposição de regras, por meio de uma chamada ao $\text{VerificaNecessidadeJitter}(PosX, PosY, L)$, detalhado no Algoritmo 2.

Algoritmo 1 $GARVis_{alg}$ – $CalculaPosicaoInicial$.

Entrada: Conjunto R de regras de associação generalizadas e medidas de avaliação M .
Saída: Lista L com posição de plotagem das regras do conjunto R
1: $L \leftarrow \emptyset$;
2: $MedidaX \leftarrow \text{MedidaPosicaoX}(M)$;
3: $MedidaY \leftarrow \text{MedidaPosicaoY}(M)$;
4: **para cada** r em R **faça**
5: $PosX \leftarrow \text{CalculaPosicaoRegraEixoX}(r, MedidaX)$;
6: $PosY \leftarrow \text{CalculaPosicaoRegraEixoY}(r, MedidaY)$;
7: $PosJ \leftarrow \text{VerificaNecessidadeJitter}(PosX, PosY, L)$;
8: $L \leftarrow L \cup PosJ$;
9: **fim-para**
10: PintaTela;

O Algoritmo 2 verifica se há a necessidade de adição de *jitter*/perturbação, dadas as posições $PosX$ e $PosY$ calculadas, visando minimizar a sobreposição de regras. Assim, para cada elemento $elem$ da lista L é verificada a existência de regras que possuam as mesmas posições $PosX$ e $PosY$. No caso de sobreposição de regras são geradas perturbações para cada um dos eixos, com valores randômicos de intervalos pré-definidos, para causar uma variação na posição da regra.

Algoritmo 2 $GARVis_{alg}$ – $VerificaNecessidadeJitter$.

Entrada: Posição $PosX$ que representa a posição da regra no eixo x, posição $PosY$ que representa a posição da regra no eixo y e L lista com a posição de plotagem das r regras.
Saída: Posição $PosJ$ da regra nos eixos x e y.
1: $tamL \leftarrow \text{Tamanho}(L)$;
2: **para cada** $elem$ em $tamL$ **faça**
3: **enquanto** ($elem.x = PosX$ e $elem.y = PosY$) **faça**
4: $PosX \leftarrow PosX + \text{GeraPerturbacao}()$;
5: $PosY \leftarrow PosY + \text{GeraPerturbacao}()$;
6: **fim-enquanto**
7: **fim-para**

Se o usuário mudar a medida ativa no eixo x e/ou y, o sistema executa tarefas semelhantes às descritas no Algoritmo 1. Dada a nova medida $MedidaX$ ou $MedidaY$ selecionada, o sistema recalcula a posição da regra (utilizando $\text{CalculaPosicaoRegraEixoX}(r, MedidaX)$ para calcular a nova posição no eixo x ou $\text{CalculaPosicaoRegraEixoY}(r, MedidaY)$ para calcular a posição no eixo y), verifica se há a necessidade de jitter (utilizando $\text{VerificaNecessidadeJitter}(PosX, PosY, L)$), atualiza a lista L e pinta a tela.

Além disso, como o usuário possui a opção de utilizar um filtro, o sistema verifica se o

usuário o ativou. No caso da ativação do filtro, é verificado o que será filtrado, podendo ser pelos valores mínimos e/ou máximos das medidas do eixo x e/ou y, e/ou pelos itens presentes no antecedente e/ou conseqüente. O usuário pode realizar a combinação desses filtros da forma que desejar. Assim, após a definição dos filtros selecionados, o sistema faz a verificação das regras que satisfazem essas condições, fornecendo-as como saída. Assim, apenas as regras que satisfazem as condições impostas pelos filtros são repintadas na tela.

Por fim, se o usuário selecionar uma regra ou uma região da tela, a regra selecionada ou as regras contidas na região selecionada são adicionadas para a lista de subconjunto de regras potencialmente interessantes (*RPI*) e pintadas para destacar que foram selecionadas. O subconjunto *RPI* é a saída do algoritmo *GARVis_{alg}* para o módulo RULEE-GARVIS.

4.5 *Estudo dirigido da aplicação da metodologia utilizando o módulo RuleE-GARVis*

Utilizando a metodologia proposta, foi realizado um estudo dirigido utilizando o módulo RULEE-GARVIS. O módulo RULEE-GARVIS permite a exploração utilizando medidas objetivas e subjetivas, suportadas por visualização. O módulo fornece recursos como: exploração via consultas SQLs; visualização de gráficos X-Y e barras utilizando medidas objetivas; filtros; visualização da taxonomia; exibição da contribuição dos itens específicos no caso da regra ser generalizada; possibilidade de análise de regras redundantes, exceção e complemento; exploração textual e gráfica com medidas subjetivas; e visualização da avaliação realizada pelo usuário.

Como conjunto de dados para o estudo dirigido foi utilizado o conjunto *groceries*⁵ com informações referentes a um mês de venda de um armazém, possuindo 9835 transações com 169 itens distintos. Para a generalização das regras de associação, foi utilizada a abordagem para pós-processamento de regras de associação APRA Carvalho (2007), que

⁵Disponível no *Projeto R para Computação Estatística* em <http://www.r-project.org/>.

generaliza as regras usando o algoritmo $APRA_{alg}$. Para gerar as regras de associação o algoritmo $APRA_{alg}$ utiliza o Apriori de Christian Borgelt⁶. Foram especificados os valores de suporte e confiança mínimos igual a 0.5 e número máximo de 5 itens por regra, sendo obtidas 2170 regras. Dado o conjunto de dados, as regras de associação e o conjunto de taxonomias de diferentes níveis elaboradas por um especialista do domínio, foi generalizado antecedente das regras, obtendo 1680 regras. Assim como mencionado na Seção 4.4.2, a generalização do antecedente ou lado esquerdo da regra, possibilita uma visão do domínio da perspectiva do relacionamento entre categorias/classes de itens e itens específicos.

Na análise objetiva, primeira etapa da metodologia proposta descrita na Seção 4.2, foi selecionado um conjunto foco formado por 459 regras com valor de Lift maior que 2. Analisando as regras (Figura 4.3) e a taxonomia (Figura 4.4) no módulo RULEE-GARVIS, com as medidas objetivas IS/Cosine, Jaccard, ϕ -coefficient, Piatetsky-Shapiro's e Kappa foram selecionadas 12 regras para o conjunto de regras potencialmente interessantes (RPI), sendo 8 generalizadas e 4 específicas. Um ponto a ser ressaltado é que essas medidas objetivas foram selecionadas por serem mais adequadas para avaliação de regras de associação generalizadas no lado esquerdo da regra (Carvalho, Rezende, & Castro, 2007).

Na etapa de avaliação das regras do conjunto RPI, segunda etapa da metodologia, o usuário especialista de domínio avaliou as regras de acordo com o seu conhecimento do domínio e objetivo no momento. O usuário pode avaliar as regras como conhecimento inesperado, útil, óbvio do domínio, prévio do especialista e/ou irrelevante. Para auxiliar a avaliação, o usuário pode visualizar a regra textualmente (Figura 4.5), as regras de origem no caso de ser uma regra generalizada, bem como exceções positivas ou negativas, complementos e regras redundantes, caso existam. Como forma de auxiliar o usuário nessa avaliação, estão disponíveis gráficos X-Y semelhante com Figura 4.3, gráficos de barras (Figura 4.6), taxonomias como Figura 4.4 e a visualização da contribuição dos itens específicos no caso de ser uma regra generalizada. Utilizando todos estes recursos, das 12 regras potencialmente interessantes, 9 regras foram avaliadas como interessantes e

⁶Disponível para download em <http://www.borgelt.net/apriori.html>.

Figura 4.3: Selecionar subconjunto de Regras Potencialmente Interessantes

3 como não interessantes por já serem de conhecimento prévio.

A etapa seguinte ao cálculo das medidas subjetivas (Etapa 3) é a análise com medidas subjetivas (Etapa 4). O usuário possui uma lista textual das regras avaliadas, separadas de acordo com a sua avaliação, sendo inesperado/útil ou óbvio do domínio/conhecimento prévio. Além disso, estão disponíveis um Gráfico X-Y, que exibe as regras coloridas de acordo com a sua avaliação (Figura 4.7), e a visualização da taxonomia. Após selecionar uma regra, o usuário pode selecionar uma ou mais medidas subjetivas, definindo os seus valores mínimos e máximos. As medidas subjetivas são conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados. Ao selecionar as medidas subjetivas, o usuário pode visualizar as regras textualmente, comentar e/ou atualizá-las como interessante e visualizá-las em gráficos de barras e X-Y com as medidas subjetivas. Assim, utilizando esses recursos, foram encontradas 3 regras interessantes nesta etapa.

Ao final, o usuário pode visualizar a avaliação das regras e as regras consideradas

Figura 4.4: Selecionar subconjunto RPI - *Taxonomia*

RuEE-GARVis

Evaluation | **Graphics** | **Taxonomy** | **Contribution of Specific Items**

Rule: [532], beef => root_vegetables

Positive Exception Rule:
[\[946\], beef & other_vegetables => root_vegetables](#)
[\[947\], beef & whole_milk => root_vegetables](#)

This rule represents:
☐ Unexpected knowledge
☐ Useful knowledge
☐ Obvious domain knowledge
☐ Specialist previous knowledge
☐ Irrelevant knowledge

Is this rule interesting?
☒ Yes ☐ No

Commentary:

Figura 4.5: Avaliar Regras Potencialmente Interessantes - Avaliar subconjunto RPI

interessantes (Figura 4.8 (b)). Também estão disponíveis gráficos de barra e pizza exibindo um panorama geral da avaliação, como a Figura 4.8 (a), que exhibe, entre outros, o número

RuIEE-GARVis

Figura 4.6: Avaliar Regras Potencialmente Interessantes - Exemplo de gráfico disponível

RuIEE-GARVis

Figura 4.7: Iniciar exploração com medidas subjetivas

de regras interessantes em relação ao número de regras específicas e generalizadas. No caso do estudo dirigido, foram encontradas 12 regras interessantes, sendo 5 específicas e 7 generalizadas.

(a) Panorama geral da avaliação

(b) Regras Interessantes

Figura 4.8: Visualização das Informações de uma Exploração Encerrada

O estudo dirigido mostrou que a aplicação da metodologia é viável, realmente auxiliando o usuário na compreensão e identificação de regras de associação generalizadas interessantes, uma vez que disponibiliza diversos recursos para exploração e avaliação das regras. A análise gráfica interativa facilitou a exploração do conjunto de regras, pois possibilitou a análise das regras com diferentes medidas e filtros em tempo real. Vale ressaltar que é necessária uma avaliação mais detalhada, com outras bases de dados e especialistas de domínio.

4.6 Considerações Finais

Visando alcançar os objetivos deste trabalho e confirmar as hipóteses, as atividades realizadas estão divididas em duas linhas. A primeira refere-se à especificação uma metodologia que visa melhorar a compreensibilidade e facilitar a identificação do conhecimento interessante em regras de associação generalizadas e a segunda objetiva desenvolver do Módulo de Gerenciamento do ambiente RULEE. Este capítulo apresentou essa

metodologia que facilita a compreensão e a identificação de regras de associação generalizadas interessantes, utilizando medidas de objetivas e subjetivas em conjunto com técnicas de visualização de informação. Essa combinação forneceu ao usuário diferentes mecanismos de avaliação, facilitando a sua participação e a descoberta de conhecimentos interessantes. Esta metodologia visa aproveitar a facilidade humana de interpretação visual e as vantagens das medidas de avaliação, fornecendo subsídios ao usuário com relação ao entendimento e à utilização do conhecimento adquirido, visando cumprir o objetivo deste trabalho.

Com base na metodologia proposta, foi desenvolvido o módulo RULEE-GARVIS. Utilizando este módulo para aplicar a metodologia, foi realizado um estudo dirigido com uma base de dados de vendas de um mês de um armazém. A análise objetiva permitiu ao usuário filtrar as regras por meio de consultas SQLs e análises gráficas. A análise gráfica interativa facilitou a exploração do conjunto de regras selecionado, pois possibilitou a análise das regras com diferentes medidas e filtros em tempo real. Selecionadas as regras potencialmente interessantes com a análise objetiva, estas regras foram apresentadas ao usuário, que avalia de acordo com os seus interesses e objetivos. Com a captação desse conhecimento, as medidas subjetivas são calculadas para posterior análise. Esta análise fornece um complemento para a análise objetiva, pois propicia a descoberta de regras interessantes antes não vistas. E assim como na análise objetiva, os gráficos se mostraram de suma importância no processo, auxiliando o usuário na exploração. Desta forma, o estudo dirigido mostrou que a aplicação da metodologia é viável, realmente auxiliando o usuário na compreensão e identificação de regras de associação generalizadas interessantes.

No próximo capítulo é apresentado a segunda linha de trabalho, que objetiva desenvolver do Módulo de Gerenciamento do ambiente RULEE. O RULEE é um ambiente para exploração e disponibilização de regras que auxilia a etapa de pós-processamento do conhecimento. Este ambiente está sendo desenvolvido no LABIC segundo a estrutura descrita na Seção 3.4.6.

Módulo de Gerenciamento do Ambiente RuEE

5.1 *Considerações Iniciais*

A existência de um ambiente para exploração de regras durante a etapa de pós-processamento e disponibilização do conhecimento é muito importante, devido à necessidade de identificação do conhecimento interessante e da participação de usuários especialistas do domínio. Assim, visando suprir esta necessidade, está sendo desenvolvido no LABIC-ICMC-USP um ambiente de exploração de regras denominado RuEE, como descrito na Seção 3.4.6.

Um fator que enfatiza a importância de ambientes e metodologias para auxiliar o usuário na etapa de pós-processamento do conhecimento é a dificuldade de interpretação devido, em grande parte, a grande quantidade de padrões gerados por muitos algoritmos. Isso ocorre com muita frequência em regras de associação, pois os algoritmos de regras de associação normalmente produzem milhares de regras, principalmente quando os atributos da base de dados são altamente correlacionados. Dessa forma, a compreensão e identifi-

cação do conhecimento interessante, em geral, tornam-se difíceis. Além disso, os trabalhos relacionados descritos no Capítulo 3, não auxiliam, ao mesmo tempo, a compreensão e a identificação do conhecimento interessante. Assim, no Capítulo 4 foi apresentado uma metodologia que utiliza medidas de avaliação objetivas e subjetivas em conjunto com técnicas de visualização de informação, e um módulo denominado RULEE-GARVIS que viabiliza a aplicação da metodologia. O módulo RULEE-GARVIS é parte integrante do ambiente RULEE, aumentando o seu poder de análise.

Também fez parte deste trabalho o projeto e a implementação do Módulo de Gerenciamento do RULEE. O Módulo de Gerenciamento possui a função de gerenciar as informações sobre os usuários do ambiente e manter os Repositórios com valores consistentes. Assim, para atingir um dos objetivos descritos na Seção 1.2, este capítulo descreve o Módulo de Gerenciamento do RULEE.

5.2 Detalhamento do Módulo de Gerenciamento

O Módulo de Gerenciamento foi projetado após ser realizada uma análise da arquitetura do RULEE definida pelo grupo, exibida na Figura 3.18, a partir do protótipo do ambiente RULEE (Paula, 2003) e dos seus módulos, *ARInE* (Melanda, 2004), RULEE-RAG (Fujimoto, Carvalho, & Rezende, 2007), RULEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006) e RULEE-SACT (Martins, 2006). Para o diagrama de casos de uso do núcleo básico do RULEE pode ser visualizado na Figura 5.1. Como pode ser observado no diagrama, o usuário pode efetuar o login no sistema, selecionar um projeto, inserir, visualizar ou consultar conjuntos de regras, além de visualizar as medidas de avaliação disponíveis.

Já na Figura 5.2 é ilustrado um esquema do núcleo básico do Módulo de Gerenciamento, que é a estrutura fundamental para o funcionamento do RULEE. Esse núcleo básico é composto pelas seguintes classes:

- *Session* - classe que representa uma sessão no ambiente. Essa classe retorna infor-

Figura 5.1: Diagrama de casos de uso do núcleo básico do RULEE

mações como o *status*, o usuário responsável, entre outros;

- *Project* - classe que representa um projeto existente no ambiente RULEE. Essa classe é utilizada para retornar informações referentes a um determinado projeto, como sua descrição, usuário responsável, seus conjuntos de regras, entre outros;
- *Ruleset* - classe que encapsula um conjunto de regras de classificação, regressão ou associação;
- *Rule* - classe que representa uma regra cadastrada no ambiente;
- *Measure* - classe que representa uma medida de avaliação;
- *Selection* - classe que representa uma consulta realizada no Repositório de Regras ou no Repositório de Medidas por um determinado usuário;
- *Comment* - classe que representa um comentário que pode ser anexado por um usuário a uma regra, um conjunto de regras ou a uma consulta;

- *Procedure* - classe que representa um procedimento que foi incorporado ao ambiente por um usuário. Classe implementada no ambiente para permitir que o usuário consiga implementar métodos personalizados, de modo a superar as limitações impostas por algumas linguagens de programação.

Figura 5.2: Esquema do núcleo básico do RULEE

O desenvolvimento do Módulo de Gerenciamento foi realizado aproveitando-se algumas funcionalidades existentes no protótipo, por exemplo, a gerência de projetos. Assim, para integrar os módulos existentes à nova arquitetura, exibida na Figura 3.18, foram realizadas as substituições das consultas SQLs por chamadas da Biblioteca de Acesso. Para um melhor entendimento da interação dos usuários com esses módulos e os relacionamentos entre as classes específicas implementadas para cada módulo em particular, foram adicionadas ao projeto do Módulo de Gerenciamento documentos referentes a esses módulos. Assim, nas Figuras 5.3, 5.5, 5.7 e 5.9 são exibidas os diagramas de casos de uso, mostrando a interação que os usuários possuem com os módulos. Já nas Figuras 5.4, 5.6 e 5.8, podem ser observadas o relacionamento entre as classes específicas implementadas para cada módulo em particular e as classes do núcleo básico. Esses relacionamentos estão sendo mostrados separadamente para cada classe apenas para facilitar a sua visualização.

Na Figura 5.3 é apresentado o diagrama de caso de uso para o módulo RULEE-RAG, módulo de exploração de regras de associação generalizadas descrito na Seção 3.4.6, e na Figura 5.4 podem ser visualizadas as classes com que a classe *RAG* se relacionam. A classe *RAG* representa uma exploração de um conjunto de regras de associação generalizadas de um determinado projeto.

Figura 5.3: Diagrama de casos de uso do módulo RULEE-RAG

Figura 5.4: Esquema do módulo RULEE-RAG

Na Figura 5.5 pode ser observado o diagrama de caso de uso para o módulo RULEE-SEAR, módulo de exploração de regras de associação com medidas subjetivas descrito na Seção 3.4.6, e a Figura 5.6 mostra as classes com que a classes *SEAR* e *Translation* se relacionam. A classe *SEAR* representa um processo de exploração subjetiva de um

conjunto de regras de associação, e a classe *Translation* representam as traduções das regras de um determinado conjunto de regras.

Figura 5.5: Diagrama de casos de uso do módulo RULEE-SEAR - contido em Sinoara, Fujimoto, & Rezende (2006)

Figura 5.6: Esquema do módulo RULEE-SEAR

Finalmente, na Figura 5.7 é exibida o diagrama de caso de uso para o módulo RULEE-

SACT, módulo de gerção semi-automática de taxonomias do ambiente RULEE descrito na Seção 3.4.6, e na Figura 5.8 as classes com que a classe *SACT* do módulo RULEE-SACT se relacionam. A classe *SACT* representa um processo de geração semi-automática de taxonomias.

Figura 5.7: Diagrama de casos de uso do módulo RULEE-SACT

Como mencionado no começo desta seção, o Módulo de Gerenciamento foi projetado e implementado levando-se em consideração os módulos desenvolvidos baseados na arquitetura do protótipo, entre eles o *ARInE*, módulo de exploração de regras de associação com medidas objetivas descrito na Seção 3.4.6. Para este módulo não foi ilustrado nenhum

Figura 5.8: Esquema do módulo RULEE-SACT

esquema, pois o *ARInE* implementa diversos métodos nas classes do núcleo básico, não possuindo uma classe exclusiva. Contudo, é importante ressaltar que esse módulo foi levado em consideração durante o projeto do Módulo de Gerenciamento. Na Figura 5.9 é mostrado o diagrama de caso de uso para o módulo *ARInE*.

Figura 5.9: Diagrama de casos de uso do módulo *ARInE*

Um outro ponto que deve ser destacado é a existência de métodos redundantes, devido ao fato de cada módulo ter sido implementado de acordo com as necessidades dos autores

em particular. Assim, observando esse problema, foi proposto criar uma interface entre as classes do núcleo básico do RULEE e os novos módulos, como pode ser observado na Figura 5.10.

Figura 5.10: Esquema da interface para os novos módulos

A configuração dos esquemas dos módulos apresentados foi mantida, excluindo apenas os métodos redundantes identificados. Assim, a partir do projeto definido nesta etapa, o primeiro passo para o desenvolvimento do Módulo de Gerenciamento foi a substituição das chamadas das consultas SQLs por chamadas da Biblioteca de Acesso. Para ser possível a realização desta atividade, foi criado um novo diretório denominado *rulee2*. Esse novo diretório foi criado com a finalidade de realizar as alterações necessárias sem a necessidade de parar o funcionamento do ambiente RULEE. Assim, a realização da substituição iniciou-se pelas classes do núcleo básico. Entretanto, uma dificuldade enfrentada pela autora foram os inúmeros problemas iniciais que consumiram mais tempo do que o desejado. Um exemplo de um problema freqüente era identificar o local e a causa exata do erro.

Na Figura 5.11 é ilustrado um exemplo de uma substituição de um código SQL por chamada da Biblioteca de Acesso. Nesta figura é mostrada parte do código do método *GetRulesetsGeneralizedProject* da classe *Project*. Assim, utilizando o método *FindPosProcGeneralizedRulesetsByProjectCode* da classe *RulesetDAO* da Biblioteca de Acesso, e realizando as modificações necessárias para que essa alteração não modifique o resultado de saída do método alterado, é gerado o código da Figura 5.12.

Essas alterações foram realizadas em todo o RULEE. No geral, foram realizadas 226 substituições no núcleo básico seguindo o mesmo padrão do exemplo anterior. Concluindo

```

    $select = $this->GetDbHandler()->prepare("select RULESET_ID FROM RULESET WHERE
PROJECT_CODE = '". $this->{_project_code}.'" AND GENERALIZED = 'YES' AND
PROCESS_STAGE='PosProc' ORDER BY RULESET_ID");

    if (!$select->execute()) {

        $err = Exception->new('EnvironmentError');
        $err->exitcode(-12);
        $err->raise("Error -12: Ruleset not found");

    } else {

        # Para cada elemento do resultado do select
        while(@array = $select->fetchrow_array()) {

            $ruleset = Ruleset->Assign($array[0], $this);

            # Insere elemento no vetor
            push(@rulesetsproject, $ruleset);

        }

    }

    $select->finish;

```

Figura 5.11: Código do protótipo do ambiente RULEE

```

$ruleset_dao = RulesetDAO->new();

@array = $ruleset_dao->FindPosProcGeneralizedRulesetsByProjectCode("$this->{_project_code}");

foreach $temp (@array) {

    $ruleset = Ruleset->Assign($temp, $this);

    # Insere elemento no vetor
    push(@rulesetsproject, $ruleset);

}

```

Figura 5.12: Código com a consulta SQL substituída pela chamada da Biblioteca de Acesso

a substituição no núcleo básico, foram realizadas 275 substituições nos módulos *ARInE*, RULEE-SEAR, RULEE-RAG e RULEE-SACT.

É importante salientar que os testes realizados durante a implementação das substituições para verificar o seu correto funcionamento utilizaram o seguinte procedimento: primeiramente, substituir as consultas SQLs nas classes necessárias para o funcionamento de uma parte do RULEE; em seguida, realizar os testes por meio da interface *Web* do ambiente; e a partir desses testes, verificar os erros. No caso deles existirem, eram identificados o local e realizadas as correções necessárias. Esse processo de verificação dos erros foi iterativa, sendo que o seu critério de parada foi a não ocorrência de erros.

Assim, a implementação do Módulo de Gerenciamento do ambiente RULEE, foi real-

izada na linguagem de programação Perl, por ser uma linguagem de programação *Web* e ser adequada para o uso de expressões regulares.

5.3 *Considerações Finais*

Neste capítulo foi apresentado o Módulo de Gerenciamento desenvolvido para o ambiente de exploração de regras RULEE. O RULEE está sendo desenvolvido com base na arquitetura exibida na Figura 3.18, apresentada na Seção 3.4.6. A arquitetura possui um conjunto de Repositórios, uma Biblioteca de Acesso, um Módulo de Gerenciamento, uma Interface, que é composta pelos módulos de Exploração e Utilitários. Atualmente, o RULEE possui implementados o conjunto de Repositórios, a Biblioteca de Acesso, o Módulo de Gerenciamento, além dos módulos RULEE-SACT (Martins, 2006), *ARInE* (Melanda, 2004), RULEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006) e RULEE-RAG (Fujimoto, Carvalho, & Rezende, 2007), apresentados na Seção 3.4.6 e do módulo RULEE-GARVIS desenvolvido neste trabalho descrito no Capítulo 4. A interface está sendo desenvolvida no momento por alunas de iniciação científica.

O Módulo de Gerenciamento do RULEE foi desenvolvido de modo a aproveitar algumas funcionalidades existentes, bem como para integrar os módulos desenvolvidos na arquitetura antiga. Para tal, foram realizadas análises nos módulos *ARInE*, RULEE-RAG, RULEE-SEAR e RULEE-SACT, para que a implementação do Módulo de Gerenciamento não comprometa o bom funcionamento dos módulos citados. A implementação do Módulo de Gerenciamento foi realizada gradualmente, sendo iniciada pela substituição das chamadas SQLs por chamadas da Biblioteca de Acesso. Ao final, foram realizados testes para a verificação do funcionamento do Módulo de Gerenciamento.

Com o desenvolvimento do Módulo de Gerenciamento, que possui a função de gerenciar as informações sobre os usuários e manter os Repositórios com valores consistentes, foi possível desenvolver o módulo RULEE-GARVIS de modo a aproveitar algumas funcionalidades oferecidas pelo ambiente RULEE. Assim, o módulo RULEE-GARVIS, descrito no Capítulo 4, utiliza funcionalidades como cálculo de medidas e inserção de regras. Tam-

bém é importante ressaltar que o desenvolvimento do RULEE-GARVIS acrescentou uma nova funcionalidade ao RULEE. O módulo RULEE-GARVIS permite a exploração e a avaliação de regras de associação generalizadas utilizando medidas de avaliação objetivas e subjetivas em conjunto com técnicas de visualização de informação.

Conclusão e Trabalhos Futuros

Este trabalho foi realizado visando confirmar a hipótese e atingir os objetivos especificados na Seção 1.1. A hipótese deste trabalho é que técnicas de visualização de informação com apoio de medidas de avaliação objetivas e subjetivas, facilitam a compreensão e a identificação do conhecimento interessante em regras de associação generalizadas. Essa combinação aproveita as vantagens que cada técnica e medida proporcionam. Como apresentado na Seção 1.2, o principal objetivo é auxiliar a compreensibilidade e facilitar a identificação de conhecimento interessante na análise de regras de associação generalizadas, por meio da utilização de técnicas de visualização de informação em conjunto com medidas objetivas e subjetivas. Para tal, foi proposta uma metodologia e implementado um módulo para viabilizar a avaliação da sua eficácia.

Também é objetivo deste projeto de pesquisa desenvolver o Módulo de Gerenciamento do ambiente para exploração de regras RULEE baseado na arquitetura descrita na Seção 3.4.6, de forma a ampliar as funcionalidades e suprir as deficiências existentes. Assim, foi desenvolvido o Módulo de Gerenciamento do ambiente RULEE.

Visando atingir os objetivos e confirmar a hipótese, nesta dissertação, inicialmente, foi realizada uma revisão bibliográfica, apresentando o processo de mineração de dados, os conceitos e definições de regras de associação e regras de associação generalizadas. Também foram discutidos algumas abordagens utilizadas para avaliação de regras, como medidas de avaliação objetivas e subjetivas, e técnicas de visualização de informação para apoiar a compreensibilidade das regras de associação. Dentre as técnicas para visualização com foco em regras de associação encontradas durante a revisão bibliográfica, apenas Yang (2005), desenvolveu uma técnica para visualização de regras de associação generalizadas. Essa técnica visa apresentar as regras utilizando coordenadas paralelas, exibindo a hierarquia dos itens da taxonomia na visualização. Entretanto, a visualização de Yang (2005) utiliza apenas as tradicionais medidas objetivas e subjetivas.

Também foram apresentados alguns trabalhos relacionados com a avaliação de regras. Dentre os trabalhos mencionados, apenas o módulo RULEE-RAG (Fujimoto, Carvalho, & Rezende, 2007) tem o objetivo de auxiliar a exploração de regras de associação generalizadas. Entretanto, o RULEE-RAG possui uma variedade de medidas disponíveis limitadas e apresenta as regras apenas na sua forma textual. Como a apresentação de muitas regras na sua forma textual pode dificultar a compreensão e a identificação de conhecimentos interessantes, foi proposta uma metodologia que utiliza técnicas de visualização de informação em conjunto com técnicas de avaliação por consulta e medidas de avaliação do conhecimento com o intuito de suprir essa necessidade. Além disso, foi desenvolvido o módulo RULEE-GARVIS, que viabiliza a aplicação da metodologia.

O módulo RULEE-GARVIS permite a exploração utilizando medidas objetivas e subjetivas, suportadas por visualização. O módulo fornece recursos como: exploração do conjunto de regras inicial via consultas SQLs ou visualização interativa de Gráfico X-Y utilizando medidas objetivas; aplicação de filtros durante a visualização de Gráfico X-Y; visualização gráfica da taxonomia, sendo que a visualização faz uma analogia com a estrutura de um diretório; exibição da contribuição dos itens específicos no caso da regra ser generalizada; avaliação das regras de acordo com o seu conhecimento e objetivo, podendo avaliar como inesperado, útil, óbvio do domínio, prévio do especialista e/ou irrelevante;

possibilidade de análise de regras redundantes, exceção e complemento; exploração textual e gráfica com medidas subjetivas; e visualização da avaliação realizada pelo usuário.

Com a realização do estudo dirigido utilizando uma base de dados de vendas de um mês de armazém, foi possível verificar que a metodologia realmente facilita a compreensão e a identificação de regras de associação generalizadas interessantes, validando a hipótese apresentada no Capítulo 1. A análise objetiva permitiu ao usuário realizar uma consulta para explorar regras que possuam algum item de prévio interesse, por exemplo, e realizar uma análise exploratória interativa com Gráfico X-Y. A análise gráfica interativa facilitou a exploração do conjunto de regras selecionado, devido a possibilidade de análise com diferentes medidas e filtros em tempo real. As regras potencialmente interessantes selecionadas com a análise objetiva foram apresentadas ao usuário, que as avaliou de acordo com os seus interesses e objetivos. A partir da captação desse conhecimento, as medidas subjetivas foram calculadas. A análise com medidas subjetivas forneceu um complemento para a análise objetiva, pois propiciou a descoberta de regras não encontradas anteriormente. Assim como na análise objetiva, os gráficos se mostraram muito importantes no processo, uma vez que auxilia o usuário durante a exploração de regras. Desta forma, o estudo dirigido mostrou que a aplicação da metodologia é viável, realmente auxiliando o usuário na compreensão e identificação de regras de associação generalizadas interessantes.

Além de viabilizar a exploração interativa de regras de associação generalizadas, o módulo RULEE-GARVIS acrescenta uma nova funcionalidade ao ambiente de exploração e disponibilização de regras RULEE. O módulo foi desenvolvido de modo a aproveitar algumas funcionalidades do RULEE, como inserção de regras e cálculo de medidas. Para tornar viável esse aproveitamento de funcionalidades no módulo RULEE-GARVIS, também foi objetivo deste trabalho auxiliar o desenvolvimento do ambiente RULEE. Este auxílio foi realizado por meio do desenvolvimento do seu Módulo de Gerenciamento, que possui a função de gerenciar as informações sobre os usuários e manter os Repositórios com valores consistentes. O Módulo de Gerenciamento do RULEE foi desenvolvido de modo a aproveitar algumas funcionalidades existentes, bem como para integrar os módulos desenvolvidos na arquitetura antiga. A implementação do Módulo de Gerencia-

mento foi realizada gradualmente, sendo iniciada pela substituição das chamadas SQLs por chamadas da Biblioteca de Acesso. Ao final, foram realizados testes para a verificação do funcionamento do Módulo de Gerenciamento.

Como o desenvolvimento da metodologia proposta, do módulo RULEE-GARVIS e do Módulo de Gerenciamento do ambiente RULEE foram atingidos os dois objetivos deste trabalho. Como principais contribuições podem ser mencionados: a metodologia que combina técnicas de visualização de informação com medidas de avaliação do conhecimento, a definição de complemento de uma regra de associação generalizada, o desenvolvimento do módulo RULEE-GARVIS e o desenvolvimento do Módulo de Gerenciamento do ambiente RULEE. Com relação ao desenvolvimento do módulo RULEE-GARVIS, podem ser destacadas: o desenvolvimento dos Gráficos interativos X-Y, no qual o usuário pode interagir com o gráfico em tempo real, “clcando” na regra, aplicando filtros ou selecionando a medida disponível desejada; o desenvolvimento da visualização da taxonomia que faz analogia com a estrutura de um diretório; possibilidade de análise de regras de complemento, exceção e redundância; disponibilidade de gráficos de barras e pizza, além dos Gráficos X-Y mencionados.

Vale destacar também que a autora desta dissertação participa do desenvolvimento do RULEE, desde a implementação dos Repositórios e Biblioteca de Acesso durante a iniciação científica; o auxílio na implementação do módulo RULEE-SEAR e RULEE-RAG. Este contato com o desenvolvimento do ambiente RULEE e de seus módulos, facilitou o desenvolvimento deste trabalho devido ao conhecimento das necessidades percebidas em experimentos com os módulos anteriormente desenvolvidos. Além disso, permitiu um melhor aproveitamento das funcionalidades existentes no RULEE. Como resultado dessa participação, alguns trabalhos foram publicados: relatório técnico do módulo RULEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006), relatório técnico do RULEE-RAG (Fujimoto, Carvalho, & Rezende, 2007) e um artigo em conferência que aborda a inserção do usuário no processo de extração de regras de associação (Yamamoto, Oliveira, Fujimoto, & Rezende, 2007). Outros foram aceitos e estão em fase de revisão: um capítulo de livro que aborda a combinação de medidas objetivas e subjetivas para identificar re-

gras interessantes (Rezende, Melanda, Fujimoto, Sinoara, & Carvalho, 2008) e um artigo para o workshop de teses e dissertações de Inteligência Artificial discutindo a metodologia proposta neste trabalho (Fujimoto & Rezende, 2008b). Outros trabalhos estão em fase de preparação: relatório técnico do módulo RULEE-GARVIS (Fujimoto & Rezende, 2008a), um artigo sobre o ambiente RULEE a ser submetido para um periódico e um artigo envolvendo a abordagem de generalização de regras de associação Carvalho (2007) e a metodologia proposta neste trabalho para exploração dessas regras a ser submetido a uma conferência internacional.

Visando trazer outras contribuições que venham a complementar este trabalho, algumas propostas para trabalho futuro são: realização de experimentos utilizando outras bases de dados e especialistas do domínio; exploração de outras técnicas de visualização de informação; exploração e incorporação de outras medidas objetivas e subjetivas; expansão da metodologia para ser utilizada em mineração de dados não-estruturados.

Referências Bibliográficas

- Adamo, J.-M. (2001). *Data Mining for Association Rules and Sequential Patterns*. Springer-Verlag. 18, 19
- Aggelis, V. (2004). Association rules model of e-banking services. In A. Zanasi, N. F. F. Ebecken, & C. A. Brebbia (Eds.), *Data Mining V*, pp. 46–55. WIT Press. 15
- Agrawal, R., T. Imielinski, & A. N. Swami (1993). Mining association rules between sets of items in large databases. In P. Buneman & S. Jajodia (Eds.), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–216. Disponível em: <http://citeseer.ist.psu.edu/agrawal93mining.html> [23/06/2008]. 17
- Agrawal, R. & R. Srikant (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, & C. Zaniolo (Eds.), *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB'94*, pp. 487–499. Disponível em: <http://citeseer.ist.psu.edu/agrawal94fast.html> [23/06/2008]. 16
- Antonie, M. & O. Zaïane (2004). Mining positive and negative association rules: An approach for confined rules. Relatório Técnico TR04-07, Dept. of Computing Science, University of Alberta. Disponível em: <http://citeseer.ist.psu.edu/antonie04mining.html> [23/06/2008]. 15
- Baesens, B., S. Viaene, & J. Vanthienen (2000). Post-processing of association rules. In

Proceedings of the special workshop on post-processing, The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2000), Boston, MA, USA, pp. 2–8. 15

Batista, G. E. A. P. A. (2003). *Pré-processamento de dados em aprendizado de máquina supervisionado*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102003-160219/> [23/06/2008]. 10

Besemann, C., A. Denton, A. Yekkirala, R. Hutchison, & M. Anderson (2004). Differential association rule mining for the study of protein-protein interaction networks. In *Proceedings of 4th Workshop on Data Mining in Bioinformatics in conjunction with The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, pp. 1–8. Disponível em: <http://citeseer.ist.psu.edu/besemann04differential.html> [23/06/2008]. 15

Blanchard, J., F. Guillet, & H. Briand (2003). Exploratory visualization for association rule rummaging. In *Proceedings of the KDD'2003 Workshop on Multimedia Data Mining MDM'03*, pp. 107–114. Disponível em: <http://citeseer.ist.psu.edu/blanchard03exploratory.html> [23/06/2008]. xix, 39, 40

Borlund, P. & P. Ingwersen (1998). Measures of relative relevance and ranked half-life: Performance indicators for interactive ir. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1998*, New York, NY, USA, pp. 324–331. ACM Press. Disponível em: <http://doi.acm.org/10.1145/290941.291019> [23/06/2008]. 26

Bruha, I. & A. Famili (2000). Postprocessing in machine learning and data mining. *ACM SIGKDD Explorations Newsletter* 2(2), pp. 110–114. Disponível em: <http://portal.acm.org/citation.cfm?doid=380995.381059> [23/06/2008]. 13

Bruzzese, D. & P. Buono (2004). Combining visual techniques for association rules

- exploration. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, New York, NY, USA, pp. 381–384. ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=989863.989930> [23/06/2008]. xix, 41
- Card, S. K., J. D. Mackinlay, & B. Shneiderman (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc. 34
- Carvalho, D. R., A. A. Freitas, & N. Ebecken (2005). Evaluating the correlation between objective rule interestingness measures and real human interest. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho, & J. Gama (Eds.), *Knowledge Discovery in Databases: Proc. of PKDD-2005. LNAI 3731*, pp. 453–461. Springer-Verlag. 2
- Carvalho, V. O. (2007). *Generalização de regras de associação utilizando conhecimento de domínio e avaliação do conhecimento generalizado*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-10122007-160534/> [23/06/2008]. 3, 21, 58, 67, 71, 75, 99
- Carvalho, V. O., S. O. Rezende, & M. Castro (2007). Evaluating generalized association rules through objective measures. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*, pp. 301–306. ACTA Press. Disponível em: <http://portal.acm.org/citation.cfm?id=1295355> [23/06/2008]. 25, 29, 63, 67, 68, 76
- Chakravarthy, S. & H. Zhang (2003). Visualization of association rules over relational dbmss. In *SAC '03: Proceedings of the 2003 ACM symposium on Applied computing*, pp. 922–926. ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=952532.952714> [23/06/2008]. xix, 40, 41, 44, 48, 52, 59
- Eugenio, B. D. & M. Glass (2004). The kappa statistic: A second look. *Computational Linguistics* 30(1), 95–101. 28

- Fayyad, U. M., G. Piatetsky-Shapiro, & P. Smyth (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM* 39(11), pp. 27–34. Disponível em: <http://citeseer.ist.psu.edu/fayyad96kdd.html> [23/06/2008]. 7, 8, 14
- Fisher, D. (1996). Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research* 4, 147–180. Disponível em: <http://citeseer.ist.psu.edu/fisher95iterative.html> [23/06/2008]. 28
- Fujimoto, M. L., V. O. Carvalho, & S. O. Rezende (2007). Descrição do módulo de exploração de regras de associação generalizadas RULEE-RAG. Relatório Técnico 296, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. 4, 44, 55, 58, 67, 69, 70, 84, 93, 96, 98
- Fujimoto, M. L. & S. O. Rezende (2008a). Descrição do módulo de visualização de regras de associação generalizadas RULEE-GARVIS. Relatório técnico, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. (Em preparação). 70, 99
- Fujimoto, M. L. & S. O. Rezende (2008b). Uma abordagem para incrementar o potencial de avaliação de regras de associação generalizadas. In *V Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence (WTDIA 2008)*. 99
- Geng, L. & H. J. Hamilton (2006). Interestingness measures for data mining: A survey. *ACM Comput. Surv.* 38(3), 9. 29
- Glymour, C., D. Madigan, D. Pregibon, & P. Smyth (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery* 1(1), pp. 11–28. Disponível em: <http://citeseer.ist.psu.edu/77634.html> [23/06/2008]. 11
- Gonçalves, E. C., I. M. B. Mendes, & A. Plastino (2005). Mining exceptions in databases. In *Advances in Artificial Intelligence*, pp. 1076–1081. Springer-Verlag. Disponível em: <http://www.springerlink.com/content/yultqngjj0pftb8w> [23/06/2008]. 65, 66
- Gonçalves, E. C. & A. Plastino (2004). Mining strong associations and exceptions in

- the stulong data set. In *Discovery Challenge. In 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2004*. Disponível em: <http://lisp.vse.cz/challenge/ecmlpkdd2004/> [23/06/2008]. 28
- Han, J. & M. Kamber (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc. 10
- Hilderman, R. J. & H. J. Hamilton (2001). Evaluation of interestingness measures for ranking discovered knowledge. In *PAKDD '01: Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, London, UK, pp. 247–259. Springer-Verlag. Disponível em: <http://citeseer.ist.psu.edu/hilderman01evaluation.html> [23/06/2008]. 2
- Hofmann, H., A. P. J. M. Siebes, & A. F. X. Wilhelm (2000). Visualizing association rules with interactive mosaic plots. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 227–235. ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=347090.347133> [23/06/2008]. xix, 38
- Jorge, A., J. Poças, & P. Azevedo (2002). Post-processing operators for browsing large sets of association rules. In *Discovery Science*, pp. 251–263. Springer-Verlag. Disponível em: <http://www.springerlink.com/content/93ba6jpucxdwkb6/> [23/06/2008]. xix, 44, 45, 46, 47, 52, 59
- Keim, D. & M. Ward (2003). Visualization. In M. Berthold & D. J. Hand (Eds.), *Intelligent Data Analysis* (1 ed.), Capítulo 11, pp. 403–443. Springer-Verlag. xix, 34, 35, 36, 37, 38
- Kohavi, R., D. Sommerfield, & J. Dougherty (1996). Data mining using $\mathcal{MLC}++$: A machine learning library in C++. In *Tools with Artificial Intelligence*, pp. 234–245. IEEE Computer Society Press. 12
- Kumar, A., Y. L. Yip, B. Smith, & P. Grenon (2006). Bridging the gap between medical and bioinformatics: An ontological case study in

- colon carcinoma. *Computers in Biology and Medicine* 36(7-8), pp. 694–711. Disponível em: <http://www.sciencedirect.com/science/article/B6T5N-4HTCW27-1/1/cbbbd488de8d9afa57f7e75edef15791> [23/06/2008]. 16
- Kurgan, L. A. & P. Musilek (2006). A survey of knowledge discovery and data mining process models. *Knowl. Eng. Rev.* 21(1), pp. 1–24. Disponível em: <http://dx.doi.org/10.1017/S0269888906000737> [23/06/2008]. 1
- Liu, B., W. Hsu, S. Chen, & Y. Ma (2000). Analyzing the subjective interestingness of association rules. *Intelligent Systems and Their Applications, IEEE* 15(5), pp. 47–55. Disponível em: <http://citeseer.ist.psu.edu/liu00analyzing.html> [23/06/2008]. xix, 2, 30, 32, 44, 46, 52, 59, 66, 68, 72
- Lourenço, F., V. Lobo, & F. Bação (2004). Binary-based similarity measures for categorical data and their application in self-organizing maps. In *XI Jornadas de Classificação e Análise de Dados, JOCLAD 2004*. 18p. Disponível em: http://www.isegi.unl.pt/ensino/docentes/fbacao/Lourenco_joclad04.pdf [23/06/2008]. 26
- Maalouf, L. & N. Mansour (2007). Mining airline data for crm strategies. In *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*, pp. 346–351. World Scientific and Engineering Academy and Society (WSEAS). 16
- Mannila, H. (1996). Data mining: Machine learning, statistics, and databases. In P. Svensson & J. C. French (Eds.), *Proceedings of the Eighth International Conference on Scientific and Statistical Database Management*, Stockholm, Sweden, pp. 1–8. IEEE Computer Society. Disponível em: <http://citeseer.ist.psu.edu/52294.html> [23/06/2008]. 8
- Martins, C. D. (2006). Geração semi-automática de taxonomias para generalização de regras de associação. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102006-104314/> [23/06/2008]. 55, 84, 93

- Melanda, E. A. (2004). *Pós-processamento de Regras de Associação*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-13012006-171753/> [23/06/2008]. 3, 24, 29, 44, 55, 56, 63, 68, 84, 93
- Melanda, E. A. & S. O. Rezende (2003). Sintaxe padrão para representar regras de associação. Relatório Técnico 206, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Disponível em: ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/reLtec/RT_206.pdf [23/06/2008]. 52
- Melanda, E. A. & S. O. Rezende (2004). Combining quality measures to identify interesting association rules. In C. Lemaître, C. A. Reyes, & J. A. Gonzales (Eds.), *IX Ibero-American Conference on Artificial Intelligence*, Volume 3315 of *Lecture Notes in Artificial Intelligence*, pp. 441–453. Springer-Verlag. Disponível em: <http://www.springerlink.com/content/tm8c5u3mmdfm2134/> [23/06/2008]. 2
- Natarajan, R. & B. Shekar (2005). A relatedness-based data-driven approach to determination of interestingness of association rules. In *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, New York, NY, USA, pp. 551–552. ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=1066677.1066803> [23/06/2008]. 2
- Oliveira, M. C. F. & H. Levkowitz (2003). From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics* 09(3), pp. 378–394. 34
- Oliveira, S., O. Zaane, & Y. Saygin (2004). Secure association rule sharing. In H. Dai, R. Srikant, & C. Zhang (Eds.), *Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD'04*, Volume 3056 of *Lecture Notes in Artificial Intelligence*, pp. 74–85. Disponível em: <http://citeseer.ist.psu.edu/706184.html> [23/06/2008].

- Omiecinski, E. (2003). Alternative interest measures for mining associations in databases. *Knowledge and Data Engineering, IEEE Transactions on* 15(1), pp. 57–69. 2
- Ong, K.-H., K.-L. Ong, W.-K. Ng, & E.-P. Lim (2002). Crystalclear: Active visualization of association rules. Disponível em: citeseer.ist.psu.edu/573510.html [23/06/2008]. xix, 39, 40
- Paula, M. F. (2003). Ambiente para exploração de regras. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. 52, 84
- Pecina, P. & P. Schlesinger (2006). Combining association measures for collocation extraction. In *ACL*, pp. 651–658. The Association for Computer Linguistics. Disponível em: <http://acl.ldc.upenn.edu/P/P06/P06-2084.pdf> [23/06/2008]. 29
- Prati, R. C., J. A. Baranauskas, & M. C. Monard (2001). Extração de informações padronizadas para a avaliação de regras induzidas por algoritmos de aprendizado de máquina simbólico. Relatório Técnico 145, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Disponível em: ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/reLtec/RT_145.pdf [23/06/2008]. 52
- Pugliesi, J. B., D. G. Dosualdo, & S. O. Rezende (2003). Sintaxe padrão para representar regras de regressão no DISCOVER. Relatório Técnico 193, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Disponível em: ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/reLtec/RT_193.pdf [23/06/2008]. 52
- Rezende, S. O. (2004). De dados a conhecimento: Evolução e desafios. Tese de Livre Docência, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. 7, 13
- Rezende, S. O., E. A. Melanda, M. L. Fujimoto, R. A. Sinoara, & V. O. Carvalho (2008). Combining data-driven and user-driven evaluation measures to identify interesting rules. In Y. Zhao, C. Zhang, & L. Cao (Eds.), *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*. Idea Group. 99

- Rezende, S. O., J. B. Pugliesi, E. A. Melanda, & M. F. Paula (2003). Mineração de dados. In S. O. Rezende (Ed.), *Sistemas Inteligentes: Fundamentos e Aplicações* (1 ed.), Capítulo 12, pp. 307–335. Manole. 8, 9
- Ribeiro, M. X., A. G. R. Balan, J. C. Felipe, A. J. M. Traina, & C. Traina (2005). Mining statistical association rules to select the most relevant medical image features. In *First International Workshop on Mining Complex Data (IEEE MCD'05)*, Houston, USA, pp. 91–98. Disponível em: http://gbdi.icmc.usp.br/publicacoes/arquivos/Ribeiro_Mining.pdf [23/06/2008]. 15
- Ribeiro, M. X., A. J. M. Traina, C. Traina, & P. Azevedo-Marques (2008). An association rule-based method to support medical image diagnosis with efficiency. *IEEE Transactions on Multimedia* 10(2), 277–285. 16
- Silberschatz, A. & A. Tuzhilin (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering* 8(6), pp. 970–974. 24, 29
- Sinoara, R. A. (2006). Avaliação subjetiva de regras de associação. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-05052006-111530> [23/06/2008]. xx, 3, 56, 57, 62, 67, 68
- Sinoara, R. A., M. L. Fujimoto, & S. O. Rezende (2006). Descrição do módulo computacional RULEE-SEAR. Relatório Técnico 271, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Disponível em: http://www.icmc.usp.br/~biblio/download/RT_271.pdf [23/06/2008]. xx, 3, 44, 55, 56, 62, 69, 70, 84, 88, 93, 98
- Sinoara, R. A. & S. O. Rezende (2006). A methodology for identifying interesting association rules by combining objective and subjective measures. *Inteligência Artificial, Revista Iberoamericana de IA* 10(32), pp. 19–27. Disponível em: <http://dialnet.unirioja.es/servlet/articulo?codigo=2229144> [23/06/2008]. 2

Smaldon, J. & A. A. Freitas (2006). A new version of the ant-miner algorithm discovering unordered rule sets. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation, GECCO 2006*, New York, NY, USA, pp. 43–50. ACM Press.

26

Srikant, R. & R. Agrawal (1997). Mining generalized association rules. *Future Generation Computer Systems* 13(2–3), pp. 161–180. Disponível em: <http://citeseer.ist.psu.edu/srikant95mining.html> [23/06/2008]. 19, 20

Tamir, R. & Y. Singer (2006). On a confidence gain measure for association rule discovery and scoring. *The VLDB Journal The International Journal on Very Large Data Bases* 15(1), pp. 40–52. 2

Tan, P.-N., V. Kumar, & J. Srivastava (2002). Selecting the right interestingness measure for association patterns. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 32–41. ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=775047.775053> [23/06/2008]. 2

Tan, P.-N., V. Kumar, & J. Srivastava (2004). Selecting the right objective measure for association analysis. *Inf. Syst.* 29(4), pp. 293–313. Disponível em: [http://dx.doi.org/10.1016/S0306-4379\(03\)00072-3](http://dx.doi.org/10.1016/S0306-4379(03)00072-3) [23/06/2008]. 2, 27

Tan, P.-N., M. Steinbach, & V. Kumar (2005). Association analysis: Basic concepts and algorithms. In *Introduction to Data Mining*, Capítulo 6, pp. 327–414. Addison-Wesley. 26, 29

Techapichetvanich, K. & A. Datta (2005). Visar : A new technique for visualizing mined association rules. In *Advanced Data Mining and Applications*, pp. 88–95. Springer-Verlag. Disponível em: <http://www.springerlink.com/content/29dmx9gj6796dmb2/> [23/06/2008]. xix, 42, 44, 48, 49, 52, 59

Weiss, S. M. & N. Indurkha (1998). *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers Inc. 8, 11

- Yamamoto, C. H., M. C. F. Oliveira, M. L. Fujimoto, & S. O. Rezende (2007). An itemset-driven cluster-oriented approach to extract compact and meaningful sets of association rules. In *Proceedings of the Sixth International Conference on Machine Learning and Applications*, pp. 87–92. Los Alamitos: IEEE Computer Society’s Conference Publishing Services. Disponível em: <http://dx.doi.org/10.1109/ICMLA.2007.17> [23/06/2008]. 2, 44, 50, 59, 98
- Yamamoto, C. H., M. C. F. Oliveira, & S. O. Rezende (2008). Including the user in the knowledge discovery loop: interactive itemset-driven rule extraction. In *SAC ’08: Proceedings of the 2008 ACM symposium on Applied computing*, pp. 1212–1217. ACM Press. Disponível em: <http://doi.acm.org/10.1145/1363686.1363967> [23/06/2008]. xix, xx, 2, 44, 50, 51, 59
- Yang, L. (2005). Pruning and visualizing generalized association rules in parallel coordinates. *IEEE Transactions on Knowledge and Data Engineering* 17(1), pp. 60–70. xix, 42, 43, 96
- Zaki, M. J. (2004). Mining non-redundant association rules. *Data Mining and Knowledge Discovery* 9(3), 223–248. Disponível em: <http://dx.doi.org/10.1023/B:DAMI.0000040429.96086.c7> [23/06/2008]. 65
- Zhang, C. & S. Zhang (2002). *Association Rule Mining: Models and Algorithms*, Volume 2307 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag. 17
- Zheng, Z., R. Kohavi, & L. Mason (2001). Real world performance of association rule algorithms. In F. Provost & R. Srikant (Eds.), *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 401–406. ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=502512.502572> [23/06/2008]. 13