Identificação de regras de associação interessantes por meio de análises com medidas objetivas e subjetivas

Roberta Akemi Sinoara

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: 10.03.2006

Assinatura:

Identificação de regras de associação interessantes por meio de análises com medidas objetivas e subjetivas[1]

Roberta Akemi Sinoara

Orientadora: ***Profa. Dra. 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
Março/2006

[1] Trabalho realizado com o apoio do CNPq.

Este documento foi preparado utilizando-se o formatador de textos LaTeX. Sua bibliografia é gerada automaticamente pelo BibTeX, utilizando o estilo Chicago.

À minha família
e à Solange Oliveira Rezende,
que também já faz parte dela.

Agradecimentos

Aos meus pais, Eliane Ribeiro e Mario Akio Sinoara, por todo amor, carinho e educação que me deram. Foi graças ao incentivo e confiança deles que cheguei até aqui. E aos meus irmãos, Fabio Kazuo Sinoara e Guilherme Hiroshi Sinoara, por agüentarem minha impaciência e pelo carinho em me ajudar a revisar esta dissertação.

À professora, orientadora e amiga Solange Rezende, minha segunda mãe, pelos muitos ensinamentos durante esses anos inesquecíveis que trabalhamos juntas. Agradeço também por todas as oportunidades que me ofereceu e pela confiança que depositou em mim desde meu primeiro ano de graduação.

Ao meu namorado Nabil Husein, pela paciência e carinho em todos os momentos, principalmente nos de mau humor. Agradeço também pelos bons conselhos de sempre e por me ensinar a encarar a vida de uma maneira diferente, "colecionando pequenas felicidades".

À amiga Lislaine A. Prando, por me ajudar a superar os desafios nesses últimos anos.

Ao novo membro do nosso grupo Anandsing Dwarkasing, pela revisão do meu abstract.

A todos os alunos e ex-alunos da Solange, entre eles, Camila Martins, Veronica Carvalho, Marina Pivato, Maria Fernanda Moura, Marcos Paula, Marcos Domingues, Valmir Marques, Tatiana Akamine, Gerson Souza e Thiago Santos, sempre dispostos a ajudar.

Às amigas Magaly Fujimoto, pela ajuda no desenvolvimento do RulEE-SEAR, e Giselle Cantador, pelo apoio nos momentos de apuro e por não se irritar com "minha pergunta de sempre".

Ao amigo Edson Melanda, pelas muitas dicas e pela ajuda imprescindível para a conclusão deste trabalho.

Aos amigos Jaqueline Pugliesi, Walter Nagai e Robson Oliveira, por me ensinarem os primeiros passos no LABIC e por todo o apoio durante a realização de meus trabalhos de iniciação científica e de mestrado.

A todos do LABIC, em especial, à professora Maria Carolina Monard e seus alunos,

sempre com valiosas contribuições e tornando nosso laboratório um ambiente muito agradável.

Ao engenheiro Fábio Fortes, por contribuir com este trabalho.

Aos funcionários do ICMC, pelo respeito e dedicação.

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

Ao CNPq, pelo apoio financeiro que permitiu a realização deste trabalho.

Resumo

A associação é uma tarefa de mineração de dados que tem sido muito utilizada em problemas reais, porém o grande número de regras de associação que podem ser geradas dificulta a identificação de conhecimento interessante aos usuários. Para apoiar a identificação de regras interessantes podem ser utilizadas medidas de avaliação de conhecimento, que normalmente são classificadas como objetivas ou subjetivas. As medidas objetivas são mais gerais, mas podem não ser suficientes por não considerarem aspectos relacionados ao usuário ou ao domínio da aplicação. Por outro lado pode haver dificuldade em se obter a subjetividade do usuário necessária para o cálculo das medidas subjetivas. Diante desse contexto, neste trabalho é proposta uma metodologia para identificação de regras de associação interessantes que combina análises com medidas objetivas e subjetivas, visando aproveitar as vantagens de cada tipo e facilitar a participação do especialista. As medidas objetivas são utilizadas para selecionar algumas regras potencialmente interessantes para serem avaliadas por um especialista. As medidas subjetivas são calculadas utilizando essas regras com as avaliações do especialista. Essas medidas subjetivas então são utilizadas para auxiliar a identificação de regras interessantes de acordo com o conhecimento obtido durante a avaliação. Para viabilizar a aplicação dessa metodologia foi desenvolvido um módulo computacional de exploração de regras de associação com medidas subjetivas, denominado RulEE-SEAR. Utilizando esse módulo e outras ferramentas já existentes foi realizado um estudo de caso com uma base de dados real sobre qualidade de vida urbana. Nesse estudo de caso o processo de identificação de regras de associação interessantes foi realizado com especialista da área e verificou-se a viabilidade da metodologia proposta.

Abstract

Association is a data mining task which has been applied in several real problems. However, due to the huge number of association rules that can be generated, it is hard for users to identify interesting knowledge. To assist users in finding interesting rules, evaluation measures can be used. Those measures are usually divided into objective and subjective. Objective measures are more general, but they can be insufficient because they do not consider user's and domain's features. On the other hand, getting users's knowledge and interest needed to calculate subjective measures can be a difficult task. In this context, a methodology to identify interesting association rules is proposed in this work. This methodology combines analysis with objective and subjective measures, aiming to use the advantages of each kind of measure and to make user's participation easier. Objective measures are used to select some potentially interesting rules for the user's evaluation. These rules and the evaluation are used to calculate subjective measures. Then, the subjective measures are used to assist the user in identifying interesting rules according to the knowledge obtained during the evaluation. To make the methodology use practicable, a computational module, named RulEE-SEAR, was developed to explore the association rules with subjective measures. Using this module and other existing tools, a case study was done. A urban life quality database was used and a specialist in this area participated in the interesting association rules identification. That case study showed that the methodology proposed is feasible.

Sumário

Resumo **ix**

Abstract **xi**

Sumário **xiii**

Lista de Figuras **xvii**

Lista de Tabelas **xix**

1 Introdução **1**

1.1 Hipótese 3

1.2 Objetivos 3

1.3 Organização 4

2 Mineração de Dados e Associação **7**

2.1 Mineração de Dados 7

2.1.1 Conceitos e Definições 8

2.1.2 O Processo de Mineração de Dados 8

2.1.3 Considerações sobre Mineração de Dados 14

2.2 Associação 15

2.2.1 Conceitos e Definições 15

2.2.2 Desafios da Aplicação de Associação 17

2.3 Considerações Finais 18

3 Avaliação de Regras e Ferramentas de Apoio **21**

3.1 Interessabilidade de Regras 22

3.1.1 Aspectos Objetivos da Interessabilidade 23

3.1.2 Aspectos Subjetivos da Interessabilidade 24

3.2 Técnicas para Auxiliar a Identificação de Regras Interessantes 26

3.3 Medidas de Avaliação de Regras de Associação 27
3.3.1 Algumas Medidas Objetivas 29
3.3.2 Algumas Medidas Subjetivas 30
3.4 Ferramentas de Apoio à Avaliação de Regras 34
3.4.1 Ambiente RulEE 34
3.4.2 Ambiente ARInE 36
3.4.3 Aplicativo AIAS 40
3.5 Considerações Finais 41

4 Metodologia para Identificação de Regras de Associação Interessantes 43
4.1 Combinação de Medidas Objetivas e Subjetivas 44
4.2 Metodologia Proposta 45
4.3 Considerações sobre a Metodologia 49
4.4 Módulo de Exploração de Regras de Associação com Medidas Subjetivas - RulEE-SEAR 50
4.4.1 Requisitos e Casos de Uso 51
4.4.2 Repositório 52
4.4.3 Classes 53
4.4.4 Cálculo das Medidas Subjetivas 53
4.4.5 Tradução de Regras 56
4.4.6 Descrição do Processo de Exploração Subjetiva no RulEE-SEAR . 58
4.5 Considerações Finais 63

5 Estudo de Caso 65
5.1 Identificação do Problema 65
5.2 Pré-Processamento 67
5.3 Extração de Padrões 67
5.4 Pós-Processamento 68
5.4.1 Análise Objetiva 68
5.4.2 Exploração Subjetiva 72
5.5 Análise dos Resultados 74
5.6 Considerações Finais 75

6 Conclusão 77

Referências Bibliográficas 81

A Documento de Requisitos do RulEE-SEAR 89
A.1 Introdução 89
A.1.1 Escopo 89
A.1.2 Definições, Acrônimos e Abreviações 90

A.2 Descrição Geral do RulEE-SEAR 91
A.2.1 Funções do Módulo 92
A.2.2 Características dos Usuários 92
A.2.3 Suposições e Dependências 92
A.3 Requisitos Específicos 93
A.3.1 Requisitos Funcionais 93
A.3.2 Requisitos de Desempenho 96

B Casos de Uso do RulEE-SEAR 97

Lista de Figuras

2.1 Etapas do processo de mineração de dados 9
2.2 Visão integrada das tarefas, etapas e tipos de dados em mineração de dados e abrangência deste trabalho 18

3.1 Arquitetura geral do ambiente RulEE 35
3.2 Processo de identificação de regras potencialmente interessantes no $\mathcal{ARInE}$ 37
3.3 Gráficos de dispersão de valores de medidas para regras de associação . . . 38
3.4 Gráficos de distribuição de freqüência de valores de medidas para regras de associação 39
3.5 Gráficos para análise de Pareto 39

4.1 Metodologia para identificação de regras de associação interessantes 46
4.2 Casos de uso do RulEE-SEAR 51
4.3 Parte do diagrama físico da base de dados do RulEE 52
4.4 Criação de um processo de exploração no RulEE-SEAR 59
4.5 Avaliação das regras no RulEE-SEAR 60
4.6 Eliminação das regras irrelevantes no RulEE-SEAR - Seleção dos itens irrelevantes 61
4.7 Eliminação das regras irrelevantes no RulEE-SEAR - Avaliação de regras possivelmente irrelevantes 61
4.8 Análise com medidas subjetivas no RulEE-SEAR - Apresentação das regras avaliadas 62
4.9 Análise com medidas subjetivas no RulEE-SEAR - Seleção da medida subjetiva 62
4.10 Análise com medidas subjetivas no RulEE-SEAR - Apresentação das regras do conjunto foco 63

5.1 Gráfico de dispersão dos valores da medida suporte 70
5.2 Gráfico de dispersão dos valores da medida confiança 70

5.3 Gráfico de dispersão dos valores da medida lift 71
5.4 Gráfico de dispersão dos valores da medida especificidade relativa 71

B.1 Casos de uso do RulEE-SEAR 97

Lista de Tabelas

2.1 Problemática de regras de associação 18

3.1 Tabela de contingência 28

4.1 Descrição das tabelas criadas para o RulEE-SEAR 53
4.2 Descrição dos principais métodos da classe SEAR 54

5.1 Descrição dos atributos da base de dados QVU 66
5.2 Tabela atributo-valor 67
5.3 Tabela de itens-transações 67
5.4 Agrupamento das medidas objetivas obtido com a análise fatorial 69
5.5 Regras obtidas com a restrição lift > 2 71
5.6 Regras obtidas com a restrição especificidade relativa $> 0{,}9$ 72
5.7 Avaliação das regras do subconjunto RPI 73
5.8 Regras interessantes encontradas na análise com medidas subjetivas 74

B.1 Especificação do caso de uso 1 (Criar Exploração Subjetiva) 98
B.2 Especificação do caso de uso 2 (Avaliar Regras) 99
B.3 Especificação do caso de uso 3 (Eliminar Regras Irrelevantes) 100
B.4 Especificação do caso de uso 4 (Calcular Medidas Subjetivas) 101
B.5 Especificação do caso de uso 5 (Analisar Regras) 102
B.6 Especificação do caso de uso 6 (Acessar Regras Interessantes) 103
B.7 Especificação do caso de uso 7 (Continuar Exploração Subjetiva) 103
B.8 Especificação do caso de uso 8 (Visualizar Exploração Subjetiva) 104
B.9 Especificação do caso de uso 9 (Definir Traduções) 105

Capítulo

1

Introdução

O desenvolvimento de tecnologias de coleta e armazenamento de dados promoveu um acúmulo de informações nas organizações a uma taxa inimaginável há alguns anos (Zuffo, 2002). A fim de conseguir uma vantagem sobre seus concorrentes, por exemplo, melhorando seu relacionamento com os clientes, essas organizações estão cada vez mais interessadas em descobrir as informações e o conhecimento contidos em suas bases de dados. Na análise desse grande volume de dados normalmente são utilizadas técnicas que compreendem a realização de consultas SQL (*Structured Query Language*), a utilização de ferramentas OLAP (*Online Analytical Processing*) ou os mecanismos de visualização de dados. Porém, tais técnicas não possibilitam que sejam expressas algumas questões importantes para tomada de decisão, como "Quais clientes gostariam de comprar o novo produto?" (Rezende, Pugliesi, Melanda, & Paula, 2003). Diante dessa 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).

Pode-se dizer que o objetivo do processo de mineração de dados é encontrar conhecimento a partir de um conjunto de dados para ser utilizado em um processo de tomada de decisão. Portanto, é importante que esse conhecimento descoberto seja compreensível e interessante aos usuários finais, a fim de fornecer suporte no processo de tomada de decisão (Rezende, Pugliesi, Melanda, & Paula, 2003). Entretanto, do ponto de vista do usuário, um dos problemas encontrados no processo de mineração de dados é a dificuldade na compreensão dos modelos extraídos. Muitas vezes, esses modelos podem ser muito complexos ou não fazer sentido para os usuários especialistas no domínio em que o processo de mineração de dados está sendo aplicado (Pazzani, 2000). A fim de facilitar a compreensão dos modelos é comum a utilização de regras como linguagem para

representação do conhecimento.

Um outro problema que tem sido identificado na utilização do conhecimento obtido é que muitos dos algoritmos de extração geram uma enorme quantidade de padrões, sendo que poucos são realmente interessantes ao usuário (Padmanabhan & Tuzhilin, 2000). Esse problema se apresenta com freqüência quando é aplicada a tarefa de mineração de dados denominada associação. 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 descobrir todas as associações existentes nas transações de uma base de dados, o que pode levar à geração de um grande número de regras, dificultando a identificação de conhecimento interessante.

Com o intuito de transpor essa dificuldade encontrada na análise dos padrões muitas pesquisas têm focado o desenvolvimento de métodos para apoiar os usuários na identificação do conhecimento interessante (Lavrač, Flach, & Zupan, 1999; Silberschatz & Tuzhilin, 1996; Jorge, Poças, & Azevedo, 2002). Uma das técnicas mais utilizadas na avaliação de regras e busca por regras interessantes é a aplicação de medidas de avaliação de conhecimento. Essas medidas normalmente são classificadas como objetivas (*data-driven*) ou subjetivas (*user-driven*), sendo que as medidas objetivas dependem exclusivamente da estrutura dos padrões e dos dados utilizados no processo de extração de conhecimento enquanto as medidas subjetivas dependem fundamentalmente do interesse e/ou necessidade dos usuários que irão utilizar o conhecimento (Silberschatz & Tuzhilin, 1996). Assim, as medidas objetivas são mais gerais e independentes do domínio em que está sendo realizado o processo de mineração de dados e dos especialistas envolvidos. Por esse mesmo motivo essas medidas podem não ser suficientes na identificação de regras interessantes, visto que os objetivos e o conhecimento dos especialistas não são considerados. Já as medidas subjetivas consideram esses fatores, porém podem existir algumas limitações em seu uso. Um especialista pode encontrar dificuldade em se expressar ao fornecer seu conhecimento e seus interesses a um sistema que utilize as medidas subjetivas. Uma boa alternativa para auxiliar a identificação de regras interessantes pode ser o uso combinado de medidas objetivas e subjetivas, aproveitando as vantagens de cada tipo.

Nesse contexto, é proposta neste trabalho uma metodologia para identificação de regras de associação interessantes que combina o uso de medidas objetivas e subjetivas. No início do processo são utilizadas as medidas objetivas, filtrando o conjunto de regras para facilitar a participação do especialista no fornecimento de seu conhecimento e seus interesses para o cálculo das medidas subjetivas. As medidas objetivas foram exploradas em um projeto anterior a este (Melanda, 2004), assim, o foco deste trabalho são as medidas subjetivas e a combinação destas com as medidas objetivas.

Neste trabalho também é apresentado o módulo de exploração de regras de associação com medidas subjetivas RuLEE-SEAR, projetado e desenvolvido com o objetivo de via-

bilizar a realização da análise com medidas subjetivas da metodologia proposta. A análise com medidas objetivas, utilizada na metodologia, é realizada no ambiente $\mathcal{ARInE}$, desenvolvido no LABIC[1] (Laboratório de Inteligência Computacional - ICMC/USP) (Melanda, 2004). Tanto o $\mathcal{ARInE}$ quanto o RulEE-SEAR são módulos de pós-processamento do ambiente RulEE (*Rule Exploration Environment*), que é um ambiente para exploração de regras de classificação, regressão e associação, cujo projeto e protótipo também foram desenvolvidos por pesquisadores do LABIC (Paula, 2003).

A hipótese que guiou o desenvolvimento deste trabalho é apresentada a seguir. Logo após, são apresentados os objetivos, gerais e específicos, e a organização desta dissertação.

1.1 Hipótese

É possível combinar a aplicação de medidas de avaliação de regras de associação objetivas e subjetivas de maneira que sejam aproveitadas as vantagens desses dois tipos de medidas e minimizadas suas desvantagens. Essa combinação pode facilitar a participação do especialista do domínio (ou do usuário final) na etapa de pós-processamento do conhecimento e permitir a identificação de regras de associação interessantes a esse especialista de maneira mais eficiente do que com o uso de apenas um tipo de medida.

1.2 Objetivos

Os objetivos do trabalho descrito nesta dissertação são apresentados a seguir, divididos em objetivos gerais e objetivos específicos.

Objetivos Gerais:

1. Propor uma metodologia combinando análises com medidas objetivas e subjetivas para auxiliar a identificação de regras de associação interessantes.

2. Verificar os benefícios do uso da análise com medidas subjetivas em conjunto com a análise com medidas objetivas na identificação de regras interessantes a determinado usuário ou problema.

3. Facilitar a participação dos usuários no processo de identificação de regras de associação interessantes.

Objetivos Específicos:

[1] http://labic.icmc.usp.br

1. Investigar algumas medidas subjetivas e o uso da subjetividade do especialista do domínio (ou usuário final) na identificação de conhecimento interessante.

2. Combinar o fluxo de análise com medidas objetivas proposto por Melanda (2004) com análises com medidas subjetivas.

3. Desenvolver um módulo computacional, integrado ao RulEE, para apoiar a avaliação de regras de associação com medidas subjetivas.

4. Fornecer ao RulEE uma interface alternativa para a disponibilização de regras de associação e para o compartilhamento de conhecimento entre usuários de um mesmo projeto.

5. Realizar um estudo de caso completo, incluindo as análises objetivas e subjetivas, com a participação de um especialista do domínio a fim de verificar a viabilidade a metodologia proposta.

1.3 Organização

Esta dissertação está dividida em 6 capítulos. Nesse primeiro capítulo foi apresentado o contexto em que se insere este trabalho, bem como a hipótese que guiou seu desenvolvimento e seus objetivos.

No Capítulo 2 é descrito o processo de mineração de dados, sendo apresentados suas etapas e os usuários que participam desse processo. A associação, uma das tarefas da mineração de dados, também é descrita nesse capítulo, visto que o foco deste trabalho está nas regras de associação. Além disso, são discutidos os desafios da aplicação do processo de mineração de dados e da associação em bases de dados reais.

No Capítulo 3 é abordada a avaliação de regras, dando ênfase à avaliação da interessabilidade. São considerados os aspectos objetivos e subjetivos dessa avaliação e apresentadas algumas técnicas utilizadas na identificação de regras interessantes, entre elas a aplicação de medidas de avaliação de conhecimento. Algumas medidas, objetivas e subjetivas, que podem ser utilizadas na avaliação de regras de associação também são apresentadas nesse capítulo, bem como três ferramentas de apoio à avaliação de regras que usam essas medidas.

No Capítulo 4 é discutido o uso combinado de medidas objetivas e subjetivas na identificação de regras de associação interessantes. Ainda nesse capítulo é descrita a metodologia para identificação de regras de associação interessantes proposta neste trabalho e o módulo computacional RulEE-SEAR desenvolvido para viabilizar a aplicação dessa metodologia.

No Capítulo 5 é apresentado um estudo de caso desenvolvido com o objetivo de validar a metodologia proposta. Nesse estudo de caso foi utilizado uma base de dados real,

contando com a participação de especialista do domínio. Todas as etapas do processo de mineração de dados realizadas são apresentadas nesse capítulo.

No Capítulo 6 são apresentadas as conclusões deste trabalho e propostas para trabalhos futuros.

Após as referências bibliográficas são apresentados, nos Apêndices A e B, o documento de requisitos e as especificações dos casos de uso do módulo computacional RuleEsear, desenvolvido neste trabalho.

Capítulo

2

Mineração de Dados e Associação

O uso de métodos manuais para análise e obtenção de conhecimento a partir de dados torna essa tarefa dispendiosa (em termos financeiros e de tempo), totalmente subjetiva e inviável quando aplicada a grandes bases de dados. Nesse contexto, técnicas especializadas de extração automática de conhecimento têm sido desenvolvidas sob o nome de *extração de conhecimento de base de dados*, geralmente referenciado na literatura como *knowledge discovery in database* (KDD), *data mining* (DM) ou *mineração de dados* (MD).

Nesse capítulo é apresentada uma visão geral do processo de mineração de dados e de uma de suas tarefas, a associação, caracterizada pela extração de regras de associação. Essa tarefa, foco deste trabalho, tem recebido grande atenção, seja na área acadêmica ou em aplicações práticas, em função de sua aplicabilidade a problemas de negócio (Antonie & Zaane, 2004; Oliveira, Zaane, & Saygin, 2004; Besemann, Denton, Yekkirala, Hutchison, & Anderson, 2004; Aggelis, 2004; Hipp, Güntzer, & Nakhaeizadeh, 2002).

2.1 Mineração de Dados

A mineração de dados é uma área multidisciplinar que incorpora técnicas utilizadas em diversas áreas como Inteligência Artificial, especialmente Aprendizado de Máquina, Base de Dados e Estatística. Por isso, as técnicas utilizadas em mineração de dados não devem ser vistas como substitutas de outras formas de análises (por exemplo, *data warehousing* e *online analytical processing*), mas como práticas para melhorar os resultados das explorações feitas com as ferramentas atualmente utilizadas.

2.1.1 Conceitos e Definições

O processo de mineração de dados, ou seja, o processo de extração de conhecimento a partir de grandes massas de dados, foi definido por Fayyad, Piatetsky-Shapiro, & Smyth (1996) como o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados. Como não se pode esperar a extração de conhecimento útil e compreensível simplesmente submetendo um conjunto de dados a uma "caixa preta" (Mannila, 1996), a mineração de dados é um processo centrado na interação entre diferentes usuários, que geralmente são divididos em três classes (Rezende, Pugliesi, Melanda, & Paula, 2003):

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. Esse usuário deve conhecer profundamente as etapas que compõem esse processo.

Usuário final Representa os usuários que utilizam o conhecimento extraído no processo de mineração de dados para apoiá-lo em tomadas de decisão. Não é necessário que esse usuário possua conhecimento profundo do domínio da aplicação.

O sucesso do processo de mineração de dados depende, em parte, dessa interação. A participação do especialista do domínio e/ou do usuário final tem grande importância na definição dos objetivos e dos dados que serão utilizados no processo de mineração de dados, bem como na avaliação do conhecimento extraído.

2.1.2 O Processo de Mineração de Dados

Na Figura 2.1 é apresentado o processo de mineração de dados. Normalmente, esse processo é iterativo, sendo seu ciclo principal formado por três grandes etapas: pré-processamento, extração de padrões e pós-processamento (Rezende, Pugliesi, Melanda, & Paula, 2003). Além dessas etapas também devem ser consideradas uma fase anterior ao ciclo, referente à identificação do problema, e outra posterior, referente à utilização do conhecimento obtido. A seguir são descritas todas as etapas do processo de mineração de dados.

Identificação do Problema

A compreensão do domínio da aplicação é um pré-requisito para a extração de conhecimento útil dos dados. Nessa fase de identificação do problema é realizado um estudo do domínio da aplicação e são definidos os objetivos e metas a serem alcançados com o

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

processo de mineração de dados. Também são selecionados os conjuntos de dados a serem utilizados na extração de conhecimento.

Além disso, o conhecimento sobre o domínio adquirido nessa fase pode fornecer subsídio para todas as etapas do processo de mineração de dados (Rezende, Pugliesi, Melanda, & Paula, 2003). Mais especificamente, na etapa de pré-processamento, esse conhecimento pode ajudar os analistas na escolha do melhor conjunto de dados para realizar a extração de padrões e na identificação de valores válidos para os atributos, de critérios de preferência entre os possíveis atributos, de restrições de relacionamentos ou de informações para geração de novos atributos. Já na etapa de extração de padrões, o conhecimento sobre o domínio pode ajudar os analistas na escolha de um critério de preferência entre os modelos gerados, no ajuste dos parâmetros do processo de indução, ou mesmo na geração de um conhecimento inicial a ser fornecido como entrada para o algoritmo de mineração, com a finalidade de aumentar a eficiência no aprendizado dos conceitos, melhorar a precisão ou a compreensão do modelo final. Na etapa de pós-processamento, o conhecimento extraído pelos algoritmos de extração de padrões deve ser avaliado. Alguns critérios de avaliação utilizam o conhecimento do especialista para saber, por exemplo, se o conhecimento extraído é interessante ao usuário (Hilderman & Hamilton, 2001; Liu & Hsu, 1996;

Piatetsky-Shapiro & Matheus, 1994).

Pré-Processamento

Geralmente os dados selecionados para o processo de mineração de dados não estão em um formato adequado para a extração de conhecimento. Além disso, durante o processo de coleta de dados podem ocorrer diversos problemas que devem ser tratados, como erros de digitação e geração de dados incorretos ou inconsistentes por sensores. A aplicação direta de alguns algoritmos de extração de padrões a todo o conjunto de dados também pode ser impossibilitada por limitações de memória ou tempo de processamento. Assim, geralmente, é necessário tratar os dados antes de realizar a etapa de extração de padrões. Diferentes atividades podem ser realizadas com essa finalidade na etapa de pré-processamento, entre elas: extração e integração, transformação, limpeza e redução de dados.

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 necessária a obtenção e a unificação desses dados, formando uma única fonte de dados.

Transformação Algumas transformações comuns que podem ser aplicadas aos dados são: resumo, transformação de tipo e normalização de atributos contínuos. O *resumo* pode ser aplicado, por exemplo, agrupando os dados sobre vendas para formar resumos diários. A *transformação de tipo* de um atributo pode ser necessária para que o algoritmo de extração de padrões possa utilizá-lo mais adequadamente como, por exemplo, converter um atributo do tipo data em um outro tipo aceito pelo algoritmo. Já a *normalização de atributos contínuos* é realizada para colocar os valores em intervalos definidos como, por exemplo, entre 0 e 1.

Limpeza Os dados podem apresentar problemas advindos do processo de coleta, como erros de digitação ou de leitura dos dados por sensores. Como o resultado da mineração de dados possivelmente será utilizado em um processo de tomada de decisão, a qualidade dos dados é um fator extremamente importante. Assim, algumas técnicas de limpeza devem ser aplicadas aos dados a fim de garantir sua qualidade. Essa limpeza pode ser realizada com ou sem a utilização do conhecimento do domínio. Utilizando-se esse conhecimento, pode-se encontrar registros com valor inválido em algum atributo, granularidade incorreta ou exemplos errôneos. Já a limpeza independente de domínio pode ser realizada seguindo uma estratégia de tratamento de atributos incompletos, com a remoção de ruído e com o tratamento de conjunto de exemplos não balanceados (Batista, Carvalho, & Monard, 2000).

Redução de Dados O número de exemplos e de atributos disponíveis para análise pode inviabilizar a utilização de algoritmos de extração de padrões. A aplicação de mé-

todos para redução de dados pode ser uma solução para esse problema. A redução pode ser feita de três maneiras (Weiss & Indurkhya, 1998): redução do número de exemplos, do número de atributos e do número de valores de um atributo. A redução do *número de exemplos* é realizada 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. A abordagem mais utilizada para redução do número de exemplos é a amostragem aleatória. A redução do *número de atributos* pode ser um modo de redução do espaço de busca pela solução. É realizada por meio da remoção de um subconjunto dos atributos existentes, de maneira que isto não tenha grande impacto na qualidade da solução final. É aconselhável a aplicação desse método com o apoio do especialista do domínio, uma vez que a remoção de um atributo potencialmente útil para o modelo pode diminuir consideravelmente a qualidade do conhecimento extraído. Já para reduzir o *número de valores de um atributo*, geralmente, são aplicados métodos de discretização ou de suavização de valores de um atributo contínuo. A discretização de um atributo consiste na substituição de um atributo contínuo por um atributo discreto, por meio do agrupamento de seus valores. Na suavização de valores, o número de valores distintos de um atributo é reduzido sem discretizá-lo. Nesse método, os valores de um determinado atributo são agrupados mas, ao contrário da discretização, cada grupo de valores é substituído por um valor numérico que o represente. Esse novo valor pode ser a média, a mediana ou mesmo os valores de borda de cada grupo.

É importante ressaltar que os objetivos do processo de mineração de dados, definidos na fase de identificação do problema, devem sempre guiar a execução das atividades de pré-processamento, de maneira que o conjunto de dados utilizado apresente as características necessárias para se atingir tais objetivos. Além disso, essas atividades devem ser realizadas criteriosamente e com o devido cuidado, garantindo que as informações presentes nos dados originais continuem presentes nas amostras geradas, para que os modelos finais representem o comportamento dos dados originais.

Extração de Padrões

A etapa de extração de padrões é direcionada ao cumprimento dos objetivos definidos na identificação do problema, com a execução de um ou mais algoritmos para extração de conhecimento. A fim de atingir seus objetivos, essa etapa deve compreender: a escolha da tarefa de mineração de dados a ser empregada, a escolha do algoritmo e a extração dos padrões propriamente dita.

A escolha da tarefa é feita de acordo com os objetivos do processo de mineração de dados, definidos na fase de identificação do problema. Diferentes tarefas podem extrair diferentes tipos de conhecimento, sendo necessário decidir já no início do processo qual o

tipo de conhecimento que o algoritmo deve extrair. As possíveis tarefas de um algoritmo de extração de padrões podem ser agrupadas em atividades preditivas e descritivas. As atividades preditivas consistem na generalização de exemplos ou experiências passadas com respostas conhecidas em um modelo, denominado preditor, capaz de identificar a classe (atributo meta) de um novo exemplo (Weiss & Indurkhya, 1998). As duas principais tarefas para predição são classificação e regressão. Já as atividades descritivas, consistem 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 freqüentes de representação de padrões destacam-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 que seja mais adequado à qualquer aplicação do processo 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. É importante ressaltar que dependendo do problema podem ser necessárias diversas execuções dos algoritmos. Por exemplo, com o objetivo de obter uma avaliação mais precisa da taxa de erro de um preditor, normalmente são utilizados métodos de reamostragem (*resampling*). A reamostragem consiste na realização de vários experimentos em que o preditor é treinado e testado com diferentes conjuntos de treinamento e testes disjuntos. Com o resultado de cada um desses experimentos, é obtida uma taxa de erro do preditor para exemplos futuros. Um dos principais métodos de reamostragem é o *k-fold cross-validation*, que subdivide o conjunto de exemplos em k partições mutuamente exclusivas e, em cada iteração do método, $k-1$ partições são fornecidas ao algoritmo de aprendizado e a partição restante é utilizada para calcular a taxa de erro da iteração. A média das taxas de erro de cada uma das k iterações é usada como a taxa de erro final, estimada pelo método.

No caso da tarefa de associação, descrita na Seção 2.2, pode-se realizar apenas uma execução do algoritmo sem comprometer a qualidade do conhecimento extraído. Segundo Zheng, Kohavi, & Mason (2001), os algoritmos para obter regras de associação devem sempre gerar a mesma solução. Além disso, os parâmetros de entrada utilizados durante a geração das regras são, geralmente, suporte e confiança mínimos, e a alteração desses valores irá apenas incluir ou excluir regras do conjunto de regras extraídas.

Como o processo de mineração de dados é iterativo, pode ser necessário que a etapa de extração de padrões seja executada diversas vezes para ajustar o conjunto de parâmetros do algoritmo de mineração, visando a obtenção de resultados mais adequados aos objetivos preestabelecidos. O conjunto de padrões extraídos nessa etapa é disponibilizado ao usuário ou incorporado a um sistema inteligente após ser analisado e/ou processado na etapa de pós-processamento.

Pós-Processamento

Na etapa de pós-processamento o conhecimento extraído pode ser simplificado, avaliado, visualizado ou simplesmente documentado para o usuário final. A análise desse conhecimento poderá determinar se o processo de extração deve ser repetido ou não. Caso o conhecimento extraído não seja de interesse para o usuário ou não esteja de acordo com os objetivos preestabelecidos, pode ser necessária a volta a etapas específicas ou uma nova realização de todo o processo, ajustando-se os parâmetros utilizados na extração de padrões ou realizando-se melhorias na seleção de dados.

Assim, a etapa de pós-processamento consiste da utilização de métodos e procedimentos, que na maioria das vezes utilizam medidas de avaliação para processar o conhecimento. Esses métodos e procedimentos podem ser agrupados nas categorias apresentadas a seguir (Bruha & Famili, 2000):

Filtragem A filtragem do conhecimento pode ser realizada por meio de mecanismos de pós-poda, para o caso de árvores de decisão, ou de truncagem, no caso de conjuntos de regras de decisão. Esses mecanismos são aplicados em situações em que os algoritmos de extração de conhecimento geram árvores de decisão com muitas folhas ou regras de decisão muito específicas, cobrindo poucos exemplos (*overfitting*). Outra forma de filtragem do conhecimento, especialmente útil para regras de associação, é o emprego de restrição de atributos ou ordenação de regras por meio de medidas de avaliação.

Interpretação e Explanação O conhecimento pode ser documentado, visualizado ou modificado de forma a torná-lo compreensível para o usuário. O conhecimento extraído também pode ser comparado ao conhecimento preexistente para a verificação de conflitos ou de conformidade e pode ser sumarizado e/ou combinado com o conhecimento prévio do domínio.

Avaliação O conhecimento descoberto deve ser avaliado a fim de verificar se o mesmo cumpre os objetivos do processo de mineração de dados. A avaliação pode ser realizada por meio de critérios como precisão, compreensibilidade e interessabilidade. A avaliação do conhecimento representado por regras é abordada mais detalhadamente no Capítulo 3.

Integração do Conhecimento A integração de conhecimento refere-se ao processo de incorporar novo conhecimento em uma base de conhecimento já existente. Isso envolve determinar como diferentes conhecimentos interagem e como o conhecimento de diferentes bases de conhecimento devem ser modificados para acomodar um ao outro (Pugliesi, 2004).

Caso seja verificado, por meio das atividades de pós-processamento, que o conhecimento cumpre os objetivos do processo de mineração de dados pode ser realizada a última fase desse processo, ou seja, a utilização do conhecimento.

Utilização do Conhecimento

Essa fase sucede o ciclo principal do processo de mineração de dados. O conhecimento extraído, depois de ser avaliado e validado na etapa de pós-processamento, é consolidado na fase de utilização do conhecimento, podendo 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.

2.1.3 Considerações sobre Mineração de Dados

A descoberta de conhecimento em bases de dados é um campo de pesquisa cujo desenvolvimento tem sido dirigido principalmente pelas necessidades sociais e econômicas. Muitos dos dados armazenados possuem informações valiosas, como tendências e padrões que podem ser usados, por exemplo, para tornar as decisões de negócios mais eficientes. Considerando que, em geral, as bases contêm uma grande quantidade de dados, 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 que deve ser considerado. Quando se realiza mineração em dados reais a etapa de pós-processamento e a disponibilização do conhecimento obtido tornam-se decisivas para o sucesso do processo. Muitos dos problemas do mundo real que se beneficiam da mineração de dados envolvem atividades de risco. Isso torna necessária uma maior confiança no conhecimento extraído, não somente em termos de precisão dos modelos, mas também em termos de sua qualidade. 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, já que 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 a associação, tarefa de mineração de dados que tem grande aplicabilidade a problemas reais.

2.2 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 e outras áreas (Antonie & Zaane, 2004; Oliveira, Zaane, & Saygin, 2004; Besemann, Denton, Yekkirala, Hutchison, & Anderson, 2004; Clementini, Felice, & Koperski, 2000; Semenova, Hegland, Graco, & Williams, 2001; Aggelis, 2004). Um exemplo do emprego da associação na área comercial pode ser observado em páginas na Internet, que apresentam sugestões do tipo "... quem comprou o produto que você procura *também* comprou estes outros produtos ...".

2.2.1 Conceitos e Definições

Na classificação usualmente empregada em mineração de dados, a associação pode ser categorizada como uma atividade de mineração de dados descritiva, como citado anteriormente. Desde a sua introdução em Agrawal, Imielinski, & Swami (1993) a associação têm recebido grande atenção e sua popularização deve-se a sua aplicabilidade a problemas de negócio e à compreensibilidade inerente às regras de associação, pois até mesmo não especialistas em mineração de dados podem compreendê-las facilmente (Hipp, Güntzer, & Nakhaeizadeh, 2002).

A idéia de extrair regras de associação surgiu da análise de dados de cestas de compras, em que são geradas regras do tipo "um cliente que compra os produtos $x_1, x_2, ..., x_n$ também irá comprar o produto y com probabilidade $c\%$". Entretanto, as regras de associação não estão restritas a análises de dependência no contexto de aplicações de varejo uma vez que elas são aplicadas com sucesso a uma ampla gama de problemas como contratos de seguros e geoprocessamento.

Uma regra de associação, representada como uma implicação na forma $LHS \Rightarrow RHS$, em que LHS e RHS são, respectivamente, o antecedente (*Left Hand Side*) e o conseqüente (*Right Hand Side*) da regra, é definida da seguinte maneira (Agrawal & Srikant, 1994):

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

A regra de associação é uma implicação[1] na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$ e $LHS \cap RHS = \varnothing$. A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações T com confiança $conf$ se em $100 * conf\%$ das transações de T em que ocorre LHS ocorre também RHS. A regra $LHS \Rightarrow RHS$ tem suporte sup se em $100 * sup\%$ das transações em T ocorre $LHS \cup RHS$.

Em regras de associação, as medidas mais empregadas são suporte e confiança, tanto na avaliação do conhecimento na etapa de pós-processamento como na seleção de *itemsets* durante o processo de geração das regras. Tais medidas são definidas a seguir:

Suporte: quantifica a incidência de um *itemset* X ou de uma regra no conjunto de dados, ou seja, indica a freqüência com que X ou com que $LHS \cup RHS$ ocorre no conjunto de dados. Da maneira como foi definido, o suporte para um *itemset* X pode ser representado por:

$$sup(X) = \frac{n(X)}{N}, \tag{2.1}$$

em que $n(X)$ é o número de transações nas quais X ocorre e N é o número total de transações consideradas. Já o suporte de uma regra $LHS \Rightarrow RHS$ pode ser representado por:

$$sup(LHS \Rightarrow RHS) = sup(LHS \cup RHS) = \frac{n(LHS \cup RHS)}{N}, \tag{2.2}$$

em que $n(LHS \cup RHS)$ é o número de transações nas quais LHS e RHS ocorrem juntos e N é o número total de transações consideradas.

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

$$conf(LHS \Rightarrow RHS) = \frac{sup(LHS \cup RHS)}{sup(LHS)} = \frac{n(LHS \cup RHS)}{n(LHS)}, \tag{2.3}$$

em que $n(LHS)$ é o número de transações nas quais LHS ocorre.

Em outras palavras, o suporte representa as freqüências dos padrões e a confiança a força da implicação, ou seja, em pelo menos $100 * conf\%$ das vezes que o antecedente (LHS) ocorrer nas transações, o conseqüente (RHS) também deve ocorrer (Zhang & Zhang, 2002).

[1] A noção de implicação aqui utilizada é diferente da noção usual de implicação em inteligência artificial. Entretanto, optou-se por essa notação por ser a utilizada por Agrawal, Imielinski, & Swami (1993).

Dadas as definições de regras de associação e das medidas suporte e confiança, o problema de obtenção de regras de associação pode ser decomposto em dois sub-problemas (Agrawal, Imielinski, & Swami, 1993):

1. Encontrar todos os *k-itemsets* (conjuntos de k itens) que possuam suporte maior ou igual ao suporte mínimo especificado pelo usuário (*sup-min*). Os *itemsets* com suporte igual ou superior a *sup-min* são definidos como *itemsets* freqüentes, os demais conjuntos são denominados de *itemsets* não-freqüentes.

2. Utilizar os *k-itemsets* freqüentes, com $k \geq 2$, para gerar as regras de associação. Para cada *itemset* freqüente $l \subseteq A$, encontrar todos os subconjuntos $\tilde{a}$ de itens de l, não vazios e diferentes de l. Para cada subconjunto $\tilde{a} \subseteq l$, gerar uma regra na forma $\tilde{a} \Rightarrow (l - \tilde{a})$ se a razão de $sup(l)$ por $sup(\tilde{a})$ for maior ou igual a confiança mínima especificada pelo usuário (*conf-min*).

   Com um *itemset* freqüente $\{a, b, c, d\}$ e um subconjunto $\{a, b\}$, por exemplo, pode-se gerar a regra $a, b \Rightarrow c, d$, desde que $conf(a, b \Rightarrow c, d) \geq$ *conf-min*, em que, $conf(a, b \Rightarrow c, d) = sup(a, b, c, d)/sup(a, b)$.

2.2.2 Desafios da Aplicação de Associação

A partir da proposição da técnica de extração de regras de associação por Agrawal, Imielinski, & Swami (1993), a maioria dos esforços em pesquisas sobre a mesma tem sido direcionada ao aprimoramento do desempenho dos algoritmos de extração de regras, em especial no sentido de propor soluções mais eficientes para o problema da identificação e contagem dos *itemsets*. Em paralelo, diversos aplicativos foram desenvolvidos incorporando os avanços propiciados por essas pesquisas. Assim, a obtenção de *itemsets* freqüentes para gerar regras de associação pode ser realizada utilizando diferentes algoritmos, como: *AIS* (Agrawal, Imielinski, & Swami, 1993), *Apriori* e *AprioriTid* (Agrawal & Srikant, 1994), *SETM* (Houtsma & Swami, 1995), *Opus* (Webb, 1995), *Direct Hashing and Pruning (DHP)* (Park, Chen, & Yu, 1997), *Dynamic Set Couting (DIC)* (Brin, Motwani, Ullman, & Tsur, 1997), *Closet* (Pei, Han, & Mao, 2000), *FP-Growth* (Wang, Tang, Han, & Liu, 2002; Han, Pei, & Yin, 2000) e *Charm* (Zaki & Hsiao, 2002).

No entanto, existe outra problemática associada às regras de associação relacionada à quantidade de regras geradas, mesmo quando a técnica é aplicada a bases de dados relativamente pequenas. Na Tabela 2.1, são apresentados resultados obtidos por meio da aplicação de algoritmos de associação a algumas bases de dados. Pode-se notar a grande quantidade de regras geradas quando se compara às quantidades usuais obtidas por outras tarefas de mineração de dados (classificação e regressão, por exemplo). Esta quantidade geralmente excede a dimensão da base de dados minerada e, principalmente, vai além da capacidade de interpretação do ser humano. Adicionalmente, o especialista do domínio ou o usuário final está interessado em uma pequena fração das regras, aquelas

que apresentam algum conhecimento útil, interessante ou inovador (Imieliński & Virmani, 1998). Assim, é necessário desenvolver metodologias e ferramentas eficientes que apóiem o usuário na identificação de regras de seu interesse durante a etapa de pós-processamento do conhecimento (Paula, 2003).

Base	**Nº de atributos**	**Nº de exemplos**	**Nº de regras**	**suporte (%)**	**confiança (%)**
Talassemia Alfa	10	339	8.827	2	50
Art A	10	18.539	2.147	6	25
Art B	1.000	1.000	18.150	6	25
Art E	100	25.000	596.626	6	25
nt60tl150ni1pa20	995	60.000	1.459.070	6	25

Tabela 2.1: Problemática de regras de associação (Melanda, 2004)

2.3 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. A Figura 2.2 fornece uma visão integrada das diferentes etapas, tarefas e principais tipos de dados do processo de mineração de dados, bem como o posicionamento deste trabalho. No contexto de mineração de dados estruturados, este trabalho insere-se no pós-processamento e na disponibilização de regras de associação.

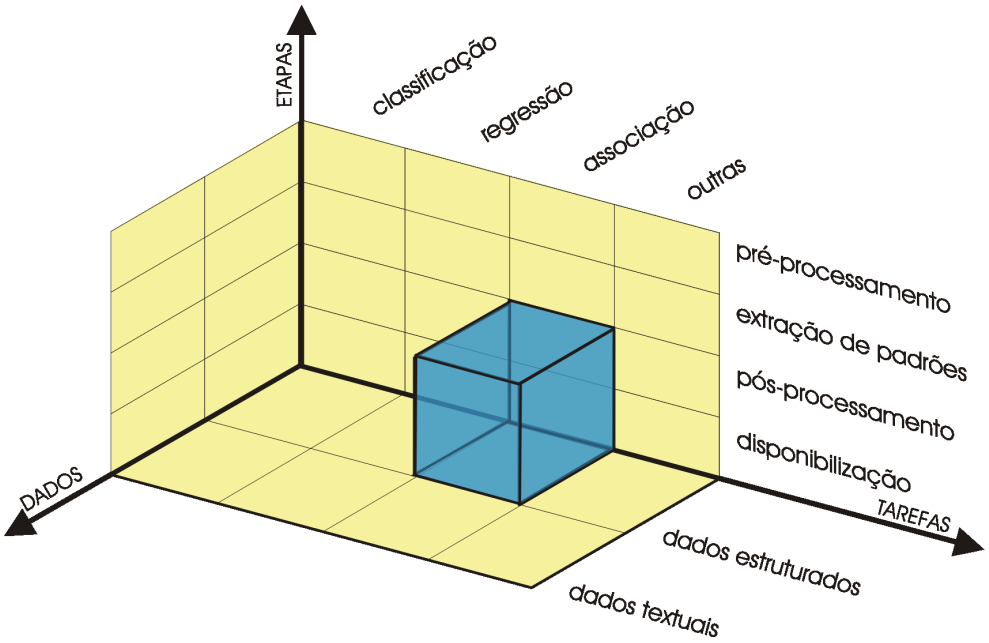


Figura 2.2: Visão integrada das tarefas, etapas e tipos de dados em mineração de dados e abrangência deste trabalho (Figura adaptada de Rezende (2004))

Quando a mineração de dados é aplicada a problemas reais deve ser dada grande

atenção à etapa de pós-processamento e à disponibilização do conhecimento, visto que deve ser garantida a qualidade do conhecimento para que ele seja utilizado em um processo de tomada de decisão. O pós-processamento do conhecimento tem especial importância para a tarefa de associação devido ao grande número de regras que podem ser geradas. Portanto, métodos para avaliação de regras de associação devem ser automatizados.

No próximo capítulo é apresentada a avaliação de regras como parte do pós-processamento do conhecimento, sendo enfatizada a avaliação de regras de associação. Também são apresentadas três ferramentas que apóiam o pós-processamento de regras, das quais duas também visam a disponibilização do conhecimento.

Capítulo

3

Avaliação de Regras e Ferramentas de Apoio

Avaliar o conhecimento é uma das atividades que podem ser realizadas durante a etapa de pós-processamento, na qual deve ser verificado se o conhecimento cumpre os objetivos definidos para o processo de mineração de dados.

Conforme apresentado na Seção 2.1, o produto final do processo de mineração de dados é um **conhecimento válido**, **novo**, **útil** e **compreensível**, que forneça aos usuários algum tipo de vantagem ao utilizá-lo. Contudo, a utilização efetiva desse conhecimento depende de fatores como a identificação do conhecimento mais interessante ou relevante e a facilidade de sua interpretação. A representação do conhecimento por meio de conjunto de regras facilita a sua interpretação, se comparado com outros modelos. No entanto, análises detalhadas são necessárias para avaliar se o conhecimento é interessante ao usuário.

Assim, os pesquisadores do processo de mineração de dados também têm dado atenção a problemas relacionados à análise, compreensão e utilização do conhecimento. Diversos trabalhos têm sido desenvolvidos e várias técnicas para avaliação de regras têm sido propostas, possibilitando avaliar tanto o desempenho quanto a qualidade das regras.

O desempenho de um conjunto de regras está associado à fidelidade com que as regras representam os dados. Geralmente, medidas como a precisão são utilizadas para quantificar o desempenho de um conjunto de regras, sendo possível identificar o **conhecimento válido**.

Já na avaliação da qualidade, são utilizados conceitos como compreensibilidade e interessabilidade (ou grau de interesse). A compreensibilidade de um dado conjunto de regras (ou de uma regra) está relacionada à sua facilidade de interpretação pelos usuários e pode ser estimada, por exemplo, pelo número de regras e/ou pelo número de condições por regra. Quanto menos condições, mais compreensível será a regra. Seguindo o mesmo raci-

ocínio para medir a compreensibilidade de um conjunto de regras: quanto menos regras, mais compreensível torna-se o conhecimento. Com essas estimativas da compreensibilidade pode-se identificar o **conhecimento compreensível**. A interessabilidade, por sua vez, é avaliada a partir de estimativas da quantidade de conhecimento interessante ao usuário no conjunto de regras ou do grau de interesse de uma regra (Hilderman & Hamilton, 2001; Tan, Kumar, & Srivastava, 2002). Dessa maneira, a interessabilidade pode estar relacionada à utilidade ou à novidade de uma regra ou de um conjunto de regras, possibilitando a identificação do **conhecimento novo e útil**.

Nesse capítulo é abordada a avaliação de regras, dando ênfase à avaliação da interessabilidade, foco deste trabalho. O problema da interessabilidade é definido, considerando seus aspectos objetivos e subjetivos, e são apresentadas algumas técnicas que auxiliam a identificação de regras interessantes. A aplicação de medidas de avaliação de conhecimento é uma dessas técnicas. Algumas medidas são apresentadas nesse capítulo, sendo que as medidas objetivas podem ser utilizadas para avaliar o desempenho e/ou a interessabilidade das regras e as subjetivas buscam apenas a identificação de regras interessantes. Por fim são apresentadas três ferramentas que apóiam a avaliação de regras: o ambiente RuleEE, o ambiente $\mathcal{ARInE}$ e o aplicativo AIAS. O RuleEE e o $\mathcal{ARInE}$, desenvolvidos no LABIC, possibilitam a avaliação de regras de maneira objetiva e o AIAS, desenvolvido por Liu, Hsu, Chen, & Ma (2000), faz uso do conhecimento do especialista durante a avaliação.

3.1 Interessabilidade de Regras

A interessabilidade refere-se ao grau de interesse de uma determinada regra descoberta para um usuário e está relacionada a fatores como novidade, utilidade, relevância e significância estatística. Visto o grande número de regras que podem ser extraídas durante a etapa de extração de padrões, principalmente quando se trata da extração de regras de associação, é inviável a análise manual da interessabilidade. Porém, para que essa análise seja automatizada é necessário que fatores de interesse sejam especificados e combinados numa medida que reflita como o especialista do domínio julga os padrões (Piatetsky-Shapiro & Matheus, 1994). A identificação desses aspectos que permitem a determinação do conhecimento interessante não é simples e depende do domínio e do usuário.

Alguns fatores que levam uma regra a não ser de interesse do usuário são (Padmanabhan & Tuzhilin, 2000; Klemettinen, Mannila, Ronkainen, Toivonen, & Verkamo, 1994):

- Uma regra pode corresponder a um conhecimento prévio.
- Uma regra pode referenciar atributos ou uma combinação de atributos não interessantes.
- Regras podem ser redundantes.

Em Liu, Hsu, Mun, & Lee (1996), a interessabilidade foi definida como: Seja B o conjunto de todos os padrões descobertos a partir de uma base de dados D pela aplicação de uma técnica Q, o problema da interessabilidade consiste em determinar o conjunto I de padrões interessantes e ordenar esses padrões de acordo com o grau de interesse do usuário em um determinado instante do tempo. Alguns aspectos importantes sobre essa definição devem ser ressaltados:

- O conjunto I pode não ser o conjunto completo de todos os padrões que foram descobertos a partir de D. Geralmente, I é apenas um subconjunto dos padrões contidos em B.
- Nem todos os padrões de I são igualmente interessantes. Dois padrões diferentes pertencentes ao conjunto I podem ter diferentes graus de interesse para o usuário.
- I é um conjunto dinâmico, pois é dependente do usuário e do tempo. Diferentes usuários podem possuir interesse em diferentes subconjuntos de B, ou ainda, um único usuário pode ter interesse em diferentes subconjuntos em diferentes momentos.

Diversas pesquisas têm sido realizadas a fim de resolver o problema da interessabilidade. Algumas dessas pesquisas visam o desenvolvimento de técnicas gerais de seleção e ordenação de regras de acordo com sua interessabilidade, considerando aspectos objetivos para essa avaliação (Freitas, 1998; Lavrač, Flach, & Zupan, 1999). Outras pesquisas consideram também aspectos subjetivos, relacionados ao usuário e ao domínio de aplicação (Silberschatz & Tuzhilin, 1996; Liu, Hsu, Mun, & Lee, 1996). Os aspectos objetivos e subjetivos da interessabilidade são apresentados a seguir.

3.1.1 Aspectos Objetivos da Interessabilidade

Considerando apenas os dados e a estrutura dos padrões, a interessabilidade pode ser estimada com a aplicação de medidas estatísticas. Vale ressaltar que somente a significância estatística não garante que uma regra é interessante.

Piatetsky-Shapiro (1991) propõe princípios intuitivos que uma medida deve satisfazer para avaliar o grau de interesse (IR) de uma regra . Segundo o autor, esse tipo de medida deve associar valores elevados a regras "fortes" e valores baixos a regras "fracas". Uma "regra forte" pode ser definida como uma descrição de regularidade, com alta confiança para uma grande quantidade de exemplos, uma "regra fraca" representa uma regularidade com alta confiança, para um número reduzido de exemplos.

Dada a regra $LHS \Rightarrow RHS$, Piatetsky-Shapiro (1991) define os princípios:

1. $RI = 0$ se $n(LHS\ RHS) = \frac{n(LHS)\ n(RHS)}{N}$. Não há interesse pela regra se o antecedente e o conseqüente são estatisticamente independentes.

2. RI aumenta uniformemente com $n(LHS\ RHS)$ enquanto outros parâmetros permanecem constantes. $RI > 0$ se LHS está correlacionado positivamente com RHS, ou seja, $n(LHS\ RHS) > \frac{n(LHS)\ n(RHS)}{N}$.

3. RI diminui uniformemente com $n(LHS)$ (ou $n(RHS)$) enquanto outros parâmetros permanecem constantes. $RI < 0$ se LHS está negativamente correlacionado com RHS, ou seja, $n(LHS\ RHS) < \frac{n(LHS)\ n(RHS)}{N}$.

Deve ser notado que esses princípios levam principalmente em consideração fatores como cobertura e confiança. Mesmo esses fatores sendo extremamente importantes na avaliação da qualidade de regras, eles não são únicos. Freitas (1999) propôs outros princípios que devem ser considerados na avaliação da interessabilidade. Alguns desses princípios são descritos a seguir:

Tamanho do disjunto Um conjunto de regras pode ser visto como uma disjunção de regras, de maneira que uma regra pode ser vista como um disjunto. O tamanho do disjunto representa o número de exemplos do conjunto de dados em que a parte condicional da regra é verdadeira, isto é, o número de exemplos cobertos pela regra. Dessa maneira, pequenos disjuntos são regras nas quais o número de exemplos cobertos é pequeno, de acordo com um critério pré-fixado. Durante a execução dos algoritmos de extração de padrões os pequenos disjuntos geralmente são eliminados dos modelos, visto que podem representar ruído presente nos dados. Entretanto, os pequenos disjuntos também podem representar uma exceção que realmente ocorre nos dados. Assim, os pequenos disjuntos devem ser tratados de maneira mais cuidadosa, evitando a eliminação de um conhecimento potencialmente interessante.

Custo dos atributos Os princípios de interesse de regras geralmente consideram o antecedente e o conseqüente da regra como um todo, sem considerar individualmente cada atributo. Entretanto, duas regras que apresentam grau de interesse parecidos considerando esses critérios podem ter avaliações muito diferentes do usuário. Em muitos domínios, os atributos podem ter diferentes "custos" para serem acessados. Um exemplo típico é o de diagnóstico médico, no qual alguns atributos podem ser determinados por exames simples, enquanto outros apenas são determinados com exames mais sofisticados. Nesse caso, o custo associado a cada atributo pode ser considerado na avaliação da regra, ou seja, quanto menor for o custo associado aos atributos presentes na regra, mais (ou menos) interessante ela é.

3.1.2 Aspectos Subjetivos da Interessabilidade

Dizer se uma regra é interessante a um usuário é uma tarefa desafiadora, uma vez que está associada a crenças, conhecimento e necessidades desse usuário. Portanto a

interessabilidade envolve estabelecer um compromisso entre esses três fatores, no qual a ponderação entre um e outro fator depende do domínio e dos objetivos estabelecidos.

Do ponto de vista do usuário, existe um conjunto de fatores que pode determinar se uma regra é interessante ou não (Hussain, Liu, & Lu, 2000). Alguns desses fatores são ortogonais entre si, outros apresentam alguma sobreposição:

Confiança A regra apresenta determinado grau de fidelidade ao ser aplicada em problemas do mundo real.

Acionamento É possível aplicar a regra de alguma maneira em um problema específico do usuário.

Utilidade A utilização da regra propicia algum benefício ao usuário, ou seja, o usuário pode obter alguma vantagem ao conhecê-la e utilizá-la.

Originalidade A regra é completamente nova para o usuário. Essa novidade não garante o acionamento ou mesmo a utilidade da regra, mas aumenta o conhecimento do usuário a respeito do domínio.

Surpresa A regra não confirma alguma expectativa ou crença do usuário.

Muitos métodos foram propostos para tentar capturar a interessabilidade de regras subjetivamente (Liu, Hsu, & Chen, 1997; Klemettinen, Mannila, Ronkainen, Toivonen, & Verkamo, 1994; Silberschatz & Tuzhilin, 1996). A idéia principal é tentar captar e utilizar o conhecimento, as experiências e os interesses dos usuários na busca por regras interessantes. Esse conhecimento têm sido utilizado tanto na etapa de extração de padrões (Wang, Jiang, & Lakshmanan, 2003) quanto no pós-processamento (Liu, Hsu, Chen, & Ma, 2000), sendo que a captação do conhecimento ou do interesse do especialista do domínio pode ocorrer de diferentes maneiras, como: especificação de uma taxonomia sobre o domínio (Shekar & Natarajan, 2004, 2002), especificação de regras em diferentes sintaxes (Zheng & Zhao, 2003; Wang et al., 2003; Liu et al., 2000; Liu & Hsu, 1996) e avaliação de algumas regras realizada diretamente pelo especialista do domínio (Sahar, 1999).

A abordagem para identificação de regras de associação interessantes proposta em Shekar & Natarajan (2004, 2002) introduz a idéia de *item-relatedness* para medir a interessabilidade entre pares de itens em uma regra de associação. Para que esse valor seja calculado, o especialista do domínio deve especificar uma taxonomia *fuzzy* que expresse sua visão sobre o domínio sendo que, nessa taxonomia, itens que possuem funcionalidades similares são representados por nós próximos. Assim, as propriedades estruturais da taxonomia são utilizadas para calcular o grau de relacionamento entre vários pares de itens de uma regra de associação. Esse grau de relacionamento entre pares de itens e a estrutura da regra são utilizados para calcular o valor da interessabilidade da regra. Esse valor é então utilizado para ordenar as regras de associação antes de apresentá-las ao usuário.

Já em uma técnica apresentada em outros trabalhos (Liu, Hsu, Mun, & Lee, 1996; Liu & Hsu, 1996), o usuário fornece o conhecimento que tem sobre o domínio e/ou sua intuição por meio de regras *fuzzy*. O conjunto de regras especificado pelo usuário é então utilizado por um algoritmo, que as compara com as regras originais. Essa técnica supõe que parte do conhecimento do domínio e do interesse do usuário está embutida nas regras especificadas pelo mesmo. Assim, as regras podem ser ordenadas de acordo com sua conformidade ou não conformidade com o conhecimento do usuário. Com esse conjunto ordenado, o usuário pode simplesmente verificar poucas regras no topo da lista para confirmar suas intuições ou conhecimento prévio ou para encontrar as regras que vão contra suas expectativas.

A avaliação realizada diretamente pelo especialista é utilizada em um processo proposto por Sahar (1999), cujo objetivo é reduzir o conjunto de regras original por meio da eliminação das regras consideradas não interessantes pelo usuário. Nesse processo as regras são agrupadas em famílias, de acordo com certas características. Logo após, algumas regras são apresentadas ao usuário, que as classifica como *"true-not-interesting"*, *"not-true-interesting"*, *"not-true-not-interesting"* ou *"true-interesting"*. A cada regra classificada pelo usuário, um conjunto de regras é eliminado de acordo com a classificação fornecida. O resultado ao final do processo é um conjunto com as regras mais interessantes.

O uso da subjetividade do usuário pode ser muito útil na busca por regras interessantes, visto que o grau de interesse de uma regra pode variar com o conhecimento e o objetivo do usuário. No entanto, a abordagem subjetiva apresenta algumas limitações:

Incerteza nos critérios Pessoas geralmente expressam seus critérios para avaliação de forma não muito precisa, por exemplo, "quanto mais anos de emprego, maior a probabilidade de um empréstimo ser aprovado".

Linguagem O usuário geralmente representa suas crenças em uma linguagem diferente dos algoritmos de aprendizado, o que dificulta a procura por padrões interessantes.

Número de regras O usuário geralmente especifica um número muito reduzido de padrões de interesse em relação ao número de regras geradas.

Busca por novos conhecimentos As crenças podem restringir a busca e limitar a descoberta de conhecimento novo.

Não há garantias das crenças do usuário As crenças do usuário podem estar erradas e restringir a busca por regras que contradizem essas crenças.

3.2 Técnicas para Auxiliar a Identificação de Regras Interessantes

Diferentes técnicas têm sido utilizadas para viabilizar a avaliação das regras obtidas na etapa de extração de padrões e, conseqüentemente, auxiliar a identificação de regras interessantes. A seguir são descritas algumas dessas técnicas (Melanda, 2004):

Aplicação de Medidas de Avaliação de Conhecimento As medidas são utilizadas com a finalidade de fornecer subsídios ao usuário no entendimento do conhecimento adquirido. Diversas medidas foram desenvolvidas para avaliar a interessabilidade das regras, sendo que algumas consideram apenas os aspectos objetivos e outras consideram também os aspectos subjetivos (Melanda, 2004; Carvalho, Freitas, & Ebecken, 2003; Liu, Hsu, Chen, & Ma, 2000; Lavrač, Flach, & Zupan, 1999).

Avaliação por Consulta Essa abordagem permite que o usuário explore o conjunto de regras por meio do uso de uma linguagem de consulta (tipicamente inspirada na SQL - *Structured Query Language*), selecionando subconjuntos de regras contendo determinados elementos e/ou satisfazendo alguma restrição definida pelo usuário (Jorge, 2004; Paula, 2003; Hipp, Mangold, Guntzer, & Nakhaeizadeh, 2002).

Poda de Regras (*Pruning*) Essa abordagem é utilizada para reduzir o número de regras geradas e tem o objetivo de eliminar (excluir) regras redundantes ou que não são interessantes para o usuário (Jorge, Poças, & Azevedo, 2002; Baesens, Viaene, & Vanthienen, 2000).

Sumarização e Generalização A técnica de sumarização tem como objetivo resumir o conhecimento extraído por meio da criação de regras gerais ou conceitos abstratos que sejam facilmente entendidos pelo usuário (Baesens, Viaene, & Vanthienen, 2000). A generalização utiliza taxonomias para transformar regras específicas em conceitos gerais, produzindo conjuntos de regras mais compactos e geralmente mais compreensíveis aos usuários (Domingues, 2004; Srikant & Agrawal, 1997).

Agrupamento de Regras (*Clustering*) O objetivo do agrupamento de regras é fornecer ao usuário grupos de regras com características semelhantes e que correspondam a diferentes temas do domínio em questão (Jorge, 2004).

As medidas de avaliação de conhecimento são descritas com mais detalhes na próxima seção. Vale ressaltar a importância das medidas na avaliação de regras, sendo freqüentemente utilizadas com outras técnicas. Por exemplo, em uma avaliação por consulta pode-se utilizar limites nos valores de medidas como restrição ao selecionar regras.

3.3 Medidas de Avaliação de Regras de Associação

Como apresentado na seção anterior, uma das abordagens utilizadas na etapa de pós-processamento do conhecimento é a aplicação de medidas para a avaliação de regras. Freqüentemente as medidas são classificadas como objetivas ou subjetivas. As medidas objetivas são mais gerais, dependendo somente da estrutura dos padrões e dos dados utilizados no processo de extração de conhecimento. Já as medidas subjetivas consideram tanto a estrutura de uma regra e os dados utilizados no processo de descoberta, como

também o conhecimento que o usuário possui e seu interesse no momento da análise dos padrões. Vale ressaltar que, apesar de considerarem a subjetividade do usuário, os valores das medidas subjetivas geralmente são obtidos de maneira objetiva, por meio de fórmulas matemáticas. Assim, também são encontrados os termos medidas *data-driven* e *user-driven* referindo-se às medidas objetivas e subjetivas, respectivamente.

Nesta seção são apresentadas algumas medidas, definidas a partir dos trabalhos de Rathjens (1996); Brin, Motwani, Ullman, & Tsur (1997); Lavrač, Flach, & Zupan (1999) e Liu, Hsu, Chen, & Ma (2000). Quanto ao objeto de avaliação, as medidas objetivas apresentadas a seguir são utilizadas para avaliar o desempenho ou a interessabilidade das regras, enquanto as medidas subjetivas visam apenas a avaliação da interessabilidade.

Visando padronizar a representação das medidas objetivas, estas são descritas com base nas freqüências associadas a cada termo da regra. As freqüências são obtidas a partir da tabela de contingência, na qual é demonstrada a cobertura da regra. Na Tabela 3.1 é apresentada a tabela de contingência para a regra $LHS \Rightarrow RHS$. Nessa tabela, LHS denota o antecedente da regra, e $\overline{LHS}$ o complemento de LHS; RHS denota o conseqüente e $\overline{RHS}$ o seu complemento; $n(LHS\ RHS)$ denota o número de transações nas quais LHS e RHS são verdadeiros; $n(\overline{LHS}\ RHS)$ denota o número de transações nas quais LHS é falso e RHS é verdadeiro; os demais elementos são representados de modo análogo; N denota o número total de transações.

	RHS	$\overline{RHS}$	
LHS	n($LHS\ RHS$)	n($LHS\ \overline{RHS}$)	n(LHS)
$\overline{LHS}$	n($\overline{LHS}\ RHS$)	n($\overline{LHS}\ \overline{RHS}$)	n($\overline{LHS}$)
	n(RHS)	n($\overline{RHS}$)	N

Tabela 3.1: Tabela de contingência

A freqüência relativa associada a qualquer elemento X da tabela de contingência pode ser calculada pela seguinte equação:

$$f(X) = \frac{n(X)}{N} \tag{3.1}$$

Assim definido, a freqüência relativa pode ser utilizada como estimativa de probabilidade. A notação $f(X|Y)$ segue a definição usual de probabilidade, dada por:

$$f(X|Y) = \frac{f(XY)}{f(Y)} \tag{3.2}$$

sendo X e Y subconjuntos das N transações.

3.3.1 Algumas Medidas Objetivas

Muitas pesquisas têm sido realizadas a fim de desenvolver medidas que não dependam do usuário ou do domínio da aplicação (Rathjens, 1996; Brin, Motwani, Ullman, & Tsur, 1997; Wang, Tay, & Liu, 1998; Lavrač, Flach, & Zupan, 1999; Adamo, 2001). Nessa seção são apresentadas algumas dessas medidas objetivas, derivadas da tabela de contingência.

Confiança esperada Quantifica a incidência de RHS no conjunto de dados, independente da ocorrência de LHS. A diferença entre confiança e confiança esperada indica o ganho de poder descritivo ao considerar LHS.

$$ConfEsp = f(RHS) \tag{3.3}$$

Lift É a razão entre confiança e confiança esperada, indicando o quanto mais freqüente se torna o RHS quando o LHS ocorre. Quanto maior o lift, maior o grau de inesperabilidade da regra.

$$Lift = \frac{f(RHS|LHS)}{f(RHS)} \tag{3.4}$$

Confiança negativa Representa a confiança na não ocorrência de associações, ou seja, dado que o antecedente é falso, o conseqüente também é falso.

$$NegRel = f(\overline{RHS} \mid \overline{LHS}) = \frac{f(\overline{LHS}\ \overline{RHS})}{f(\overline{LHS})} \tag{3.5}$$

Confiança relativa Representa o ganho de confiança em relação à regra $\mathsf{true} \Rightarrow RHS$ pela inclusão de LHS.

$$\begin{aligned} RConf = & \ f(RHS|LHS) - f(RHS) \\ = & \ \tfrac{f(LHS\ RHS)}{f(LHS)} - f(RHS) \end{aligned} \tag{3.6}$$

Sensibilidade É definida como a freqüência condicional de que LHS é verdadeiro dado que RHS é verdadeiro. Pode ser interpretada também como sendo a confiança da regra $RHS \Rightarrow LHS$.

$$Sens = f(LHS \mid RHS) = \frac{f(LHS\ RHS)}{f(RHS)} \tag{3.7}$$

Sensibilidade relativa Representa o ganho de sensibilidade em relação a $LHS \Rightarrow$ true.

$$\begin{aligned} RSens = & \; f(LHS|RHS) - f(LHS) \\ = & \; \frac{f(LHS\ RHS)}{f(RHS)} - f(LHS) \end{aligned} \tag{3.8}$$

Especificidade É dada pela freqüência condicional de que LHS seja falso dado que RHS é falso.

$$Spec = f(\overline{LHS} \mid \overline{RHS}) = \frac{f(\overline{LHS}\ \overline{RHS})}{f(\overline{RHS})} \tag{3.9}$$

Especificidade relativa De forma análoga à sensibilidade relativa, representa o ganho de especificidade em relação a $LHS \Rightarrow$ true.

$$\begin{aligned} RSpec = & \; f(\overline{RHS}|\overline{LHS}) - f(\overline{LHS}) \\ = & \; \frac{f(\overline{LHS}\ \overline{RHS})}{f(\overline{RHS})} - f(\overline{LHS}) \end{aligned} \tag{3.10}$$

Novidade Visa identificar o quão inovadora, interessante ou não-usual é uma dada regra. É definida pela comparação entre o valor observado da ocorrência de LHS e RHS e o valor esperado de ocorrência se LHS e RHS fossem independentes. Cabe destacar que o primeiro termo da medida, $f(LHS\ RHS)$, é o suporte da regra, indicando que valores elevados de suporte apenas são interessantes quando valores de $f(LHS)$ e/ou $f(RHS)$ são relativamente pequenos.

$$Nov = f(LHS\ RHS) - f(RHS) \cdot f(LHS) \tag{3.11}$$

Convicção De modo similar à medida confiança, a convicção representa o "poder" associativo entre o antecedente e o conseqüente de uma regra. Experimentos realizados por Brin, Motwani, Ullman, & Tsur (1997) indicam que as regras mais interessantes possuem o valor da medida convicção entre 1,01 e 5.

$$Conv = \frac{f(LHS) \cdot f(\overline{RHS})}{f(LHS\ \overline{RHS})} \tag{3.12}$$

Uma síntese de outras medidas objetivas derivadas da tabela de contingência são apresentadas em Melanda (2004).

3.3.2 Algumas Medidas Subjetivas

Considerando aspectos subjetivos na avaliação da interessabilidade de regras, ou seja, considerando o interesse do usuário no momento da avaliação, em Silberschatz & Tuzhi-

lin (1996) foi proposta uma classificação para as medidas subjetivas. Essa classificação identifica as duas principais razões que tornam o conhecimento interessante:

Inesperabilidade (*Unexpectedness*) O conhecimento é interessante se ele surpreende o usuário, ou seja, ele 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. Portanto, regras interessantes podem apresentá-los de maneira combinada, podendo ser: inesperadas e úteis, inesperadas e não úteis ou esperadas e úteis. Embora a inesperabilidade e a utilidade sejam importantes, a utilidade é a principal característica na maioria das aplicações, visto que regras úteis permitem que o usuário desempenhe melhor o seu trabalho (Silberschatz & Tuzhilin, 1996; Liu, Hsu, Chen, & Ma, 2000). No entanto a utilidade das regras é mais difícil de ser avaliada, pois envolve classificar regras e obter ações que devem ser tomadas para cada classe, sendo que essas ações podem variar ao longo do tempo. Assim, as pesquisas realizadas nessa área se concentram mais na avaliação da inesperabilidade e esperam atingir a utilidade, acreditando que a maioria dos regras úteis são também inesperadas e vice-versa.

Liu, Hsu, Chen, & Ma (2000) propõem quatro medidas para identificar regras de associação esperadas e inesperadas considerando o conhecimento prévio do domínio. Tais medidas são utilizadas no aplicativo AIAS, apresentado na Seção 3.4.3. Nesse aplicativo, o conhecimento prévio do domínio pode ser fornecido pelo usuário por meio da especificação de um *conhecimento impreciso* ou de uma *impressão geral*[1]. As medidas, apresentadas a seguir, são definidas em relação ao grau com que o antecedente (LHS) e/ou o conseqüente (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. Idem para R_{ij} em relação a RHS. Os valores destes fatores variam de 0 (para nenhuma conformidade) a 1 (para conformidade completa).

Conformidade Permite identificar e classificar regras em conformidade com uma impressão geral ou um conhecimento impreciso fornecido pelo usuário especialista do domínio.

$$confm_{ij} = L_{ij} \cdot R_{ij} \tag{3.13}$$

[1]Conhecimento impreciso e impressão geral são definidos na Seção 3.4.3

Antecedente inesperado Permite avaliar se o antecedente (LHS) da regra é inesperado.

$$unexpCond_{ij} = \begin{cases} 0 & \text{se } R_{ij} - L_{ij} \leq 0, \\ R_{ij} - L_{ij} & \text{se } R_{ij} - L_{ij} > 0. \end{cases} \tag{3.14}$$

Conseqüente inesperado Permite avaliar se o conseqüente (RHS) da regra é inesperado.

$$unexpConseq_{ij} = \begin{cases} 0 & \text{se } L_{ij} - R_{ij} \leq 0, \\ L_{ij} - R_{ij} & \text{se } L_{ij} - R_{ij} > 0. \end{cases} \tag{3.15}$$

Antecedente e conseqüente inesperados Permite avaliar se o antecedente e o conseqüente da regra são inesperados.

$$bsUnexp_{ij} = 1 - max\big[confm_{ij}, unexpConseq_{ij}, unexpCond_{ij}\big] \tag{3.16}$$

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

$$L_{ij} = min\left(\frac{LM_{ij}}{LN_i}, \frac{SM_{ij}}{SN_j}\right) \tag{3.17}$$

$$R_{ij} = \frac{RM_{ij}}{RN_i} \tag{3.18}$$

senão

$$R_{ij} = min\left(\frac{RM_{ij}}{RN_i}, \frac{SM_{ij}}{SN_j}\right) \tag{3.19}$$

$$L_{ij} = \frac{LM_{ij}}{LN_i} \tag{3.20}$$

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 elementos[2] da impressão geral especificada pelo usuário. 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;

- SM_{ij} se refere 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 3.21 e 3.22:

$$L_{ij} = min\left(\frac{LM_{ij}}{LN_i}, \frac{LSM_{ij}}{LSN_j}\right) \tag{3.21}$$

$$R_{ij} = min\left(\frac{RM_{ij}}{RN_i}, \frac{RVM_{ij}}{RVN_j}\right) \tag{3.22}$$

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.

Na próxima seção são apresentadas três ferramentas que apóiam a avaliação de regras, sendo que as duas primeiras utilizam as medidas objetivas apresentadas na Seção 3.3.1 e a terceira utiliza as medidas subjetivas apresentadas nessa seção.

[2]Conforme 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.

3.4 Ferramentas de Apoio à Avaliação de Regras

Diferentes ferramentas têm sido desenvolvidas para apoiar a avaliação de regras (Melanda, 2004; Paula, 2003; Jorge, Poças, & Azevedo, 2002; Liu, Hsu, Chen, & Ma, 2000; Ma, Liu, & Wong, 2000; Ma, Wong, & Liu, 2000). Nessa seção são apresentados dois ambientes projetados para apoiar a disponibilização do conhecimento representado por regras e sua avaliação por meio de medidas objetivas: o RulEE e o $\mathcal{ARInE}$. Também é apresentado o aplicativo AIAS, que busca identificar regras de associação interessantes utilizando o conhecimento prévio do domínio e medidas subjetivas.

O ambiente RulEE apóia a exploração de regras de classificação, regressão e associação por meio de filtros com medidas objetivas. Além disso ele possui módulos de pós-processamento de regras de associação. O $\mathcal{ARInE}$ é um desses módulos e possui funcionalidades para auxiliar a análise de medidas objetivas e a avaliação de regras de associação com essas medidas. Outro módulo de pós-processamento do RulEE é o módulo de exploração de regras de associação com medidas subjetivas, desenvolvido neste trabalho e apresentado na Seção 4.4.

Assim, essas três ferramentas estão relacionadas com este trabalho. A análise objetiva disponibilizada no $\mathcal{ARInE}$ e as medidas subjetivas encontradas no AIAS são utilizadas na metodologia para identificação de regras de associação interessantes que é proposta neste trabalho. Já o módulo computacional que apóia a análise com medidas subjetivas, desenvolvido para viabilizar o uso dessa metodologia, foi integrado ao RulEE como um de seus módulos de pós-processamento.

3.4.1 Ambiente RulEE

O ambiente computacional para exploração de regras denominado RulEE (*Rule Exploration Environment*) foi desenvolvido para apoiar a etapa de pós-processamento e a disponiblização do conhecimento no processo de mineração de dados Paula (2003). Esse ambiente foi projetado visando integrar diferentes técnicas e medidas para avaliação de regras em um único sistema, em que diversos tipos de análises pudessem ser feitas. A fim de verificar a utilidade desse ambiente foi desenvolvido um protótipo e, com base nesse protótipo, o ambiente RulEE está sendo desenvolvido no Laboratório de Inteligência Computacional (LABIC - ICMC/USP), contendo módulos de pós-processamento com funcionalidades adicionais não implementadas no protótipo.

Uma característica importante do RulEE é facilitar o acesso ao conhecimento descoberto. Isso é realizado por meio de uma interface para disponibilização do conhecimento baseada na Web. Dessa maneira, algumas características importantes da Web 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 de co-

nhecimento utilizada.

A arquitetura do ambiente RULEE que está em desenvolvimento é apresentada na Figura 3.1, sendo composta por:

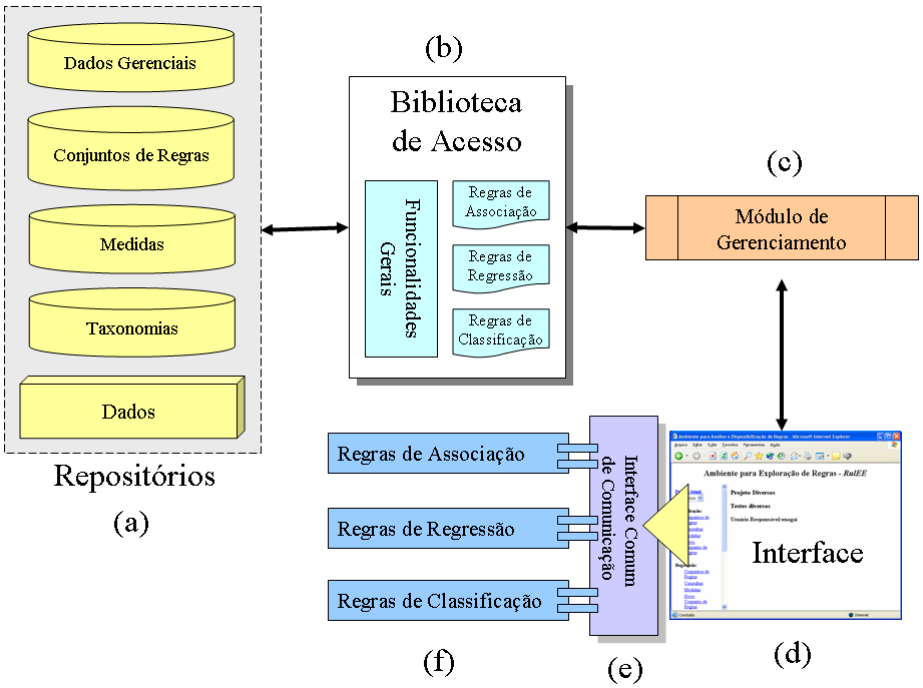


Figura 3.1: Arquitetura geral do ambiente RULEE

Repositório de Dados Gerenciais Armazena informações sobre os usuários e os projetos.

Repositório de Conjuntos de Regras Armazena conjuntos de regras de associação, classificação e regressão.

Repositório de Medidas Contém informações sobre medidas de avaliação de conhecimento, tanto objetivas quanto subjetivas. Nesse repositório também são armazenados valores das medidas calculadas para cada regra presente no repositório de conjunto de regras.

Repositório de Taxonomias Armazena taxonomias de determinados domínios. Essas taxonomias são utilizadas para generalizar regras de associação e para analisar conjuntos de regras generalizadas.

Repositório de Dados Formado pelos conjuntos de dados utilizados na extração e teste de determinado conjunto de regras.

Biblioteca de Acesso Formada por um conjunto de programas com funções específicas de acesso, manipulação e obtenção dos dados dos repositórios. A biblioteca faz a ligação entre o módulo de gerenciamento e os repositórios.

Módulo de Gerenciamento Possui a função de gerenciar as informações sobre os usuários do RulEE, manter os repositórios com valores consistentes e manter a comunicação entre os dados dos repositórios e a interface, por meio da biblioteca de acesso.

Interface 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.

O RulEE auxilia a análise de regras de classificação, regressão e associação, recebendo como entrada arquivos com conjuntos de regras na sintaxe padrão definida no Discover (Prati, Baranauskas, & Monard, 2001; Pugliesi, Dosualdo, & Rezende, 2003; Melanda & Rezende, 2003), ambiente desenvolvido no LABIC para apoiar todo o processo de extração de conhecimento. Por utilizar as regras na sintaxe padrão do Discover e com a utilização dos *scripts* de conversão dos diversos tipos de regras para a sintaxe padrão, também implementados no Discover, podem ser inseridas no RulEE 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*.

Uma vez que o conhecimento descoberto tenha sido armazenado no ambiente, as regras são disponibilizadas ao usuário juntamente com métodos para que ele possa avaliá-las. O RulEE apresenta um conjunto de medidas objetivas de avaliação que pode ser utilizado pelo usuário para ordenação e seleção de regras, auxiliando a identificação do conhecimento interessante. O ambiente permite que as medidas objetivas sejam utilizadas de forma combinada na análise das regras, por meio de consultas com a aplicação de restrições (ou filtros).

Além disso, esse ambiente foi modelado para permitir, com pequenas alterações no seu código fonte, a incorporação de novas medidas de avaliação e novas funcionalidades por meio de módulos de pós-processamento. Assim, com a utilização do RulEE os pesquisadores podem realizar experimentos com novas medidas de avaliação e novas técnicas de pós-processamento sem despender muito tempo com implementações.

Aproveitando as facilidades de integração de novas funcionalidades ao RulEE foram desenvolvidos três módulos de pós-processamento. O *ARInE*, descrito na próxima seção, o RulEE-GAR, que realiza a generalização de conjuntos de regras de associação e a análise de regras generalizadas (Domingues, 2004) e o módulo de exploração de regras de associação com medidas subjetivas (RulEE-SEAR), desenvolvido neste trabalho e apresentado no próximo capítulo.

3.4.2 Ambiente *ARInE*

O *ARInE* (*Environment for Association Rules Interactive Exploration*) foi projetado e implementado para disponibilizar um ambiente interativo com o objetivo de apoiar a

avaliação de regras de associação com medidas objetivas. Esse ambiente está integrado ao protótipo do ambiente RULEE, apresentado na Seção 3.4.1, utilizando seus repositórios de dados gerenciais, de conjuntos de regras e de medidas. Assim, o $\mathcal{ARInE}$ pode ser visto como um módulo de pós-processamento para regras de associação do RULEE.

Todas as funcionalidades do $\mathcal{ARInE}$ podem ser acessadas por sua interface Web. A análise das regras nesse ambiente segue o fluxo apresentado na Figura 3.2. Esse fluxo e as principais funcionalidades do $\mathcal{ARInE}$ são definidos por uma metodologia para análise de regras de associação fundamentada em análises gráficas de medidas objetivas (Melanda, 2004). Em linhas gerais, o objetivo dessa metodologia é a identificação de um pequeno subconjunto de regras potencialmente interessantes (RPI) segundo uma ou mais medidas objetivas para ser apresentado a um especialista ou usuário.

Conforme a Figura 3.2, a análise das regras no $\mathcal{ARInE}$ inicia-se com a definição de um conjunto de regras foco (*focus-ruleset*). O usuário define seu conjunto de regras foco realizando consultas ao conjunto de regras resultante da etapa de extração de padrões disponibilizado no RULEE. 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. Essa etapa de definição do conjunto foco é opcional, contudo é bastante útil quando o usuário está interessado em descobrir regras relacionadas a apenas alguns itens.

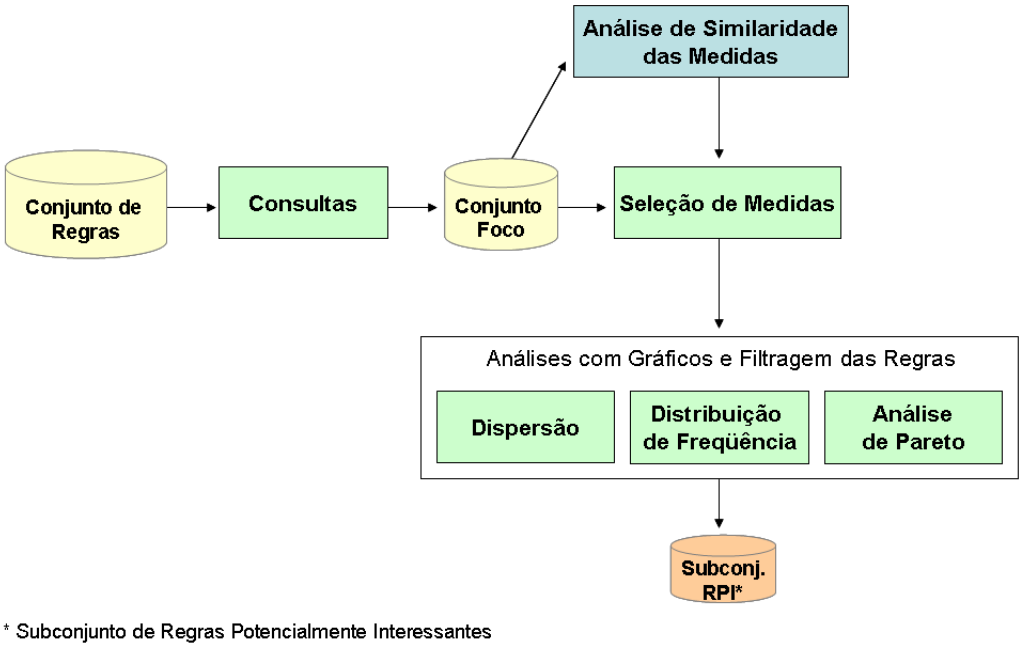
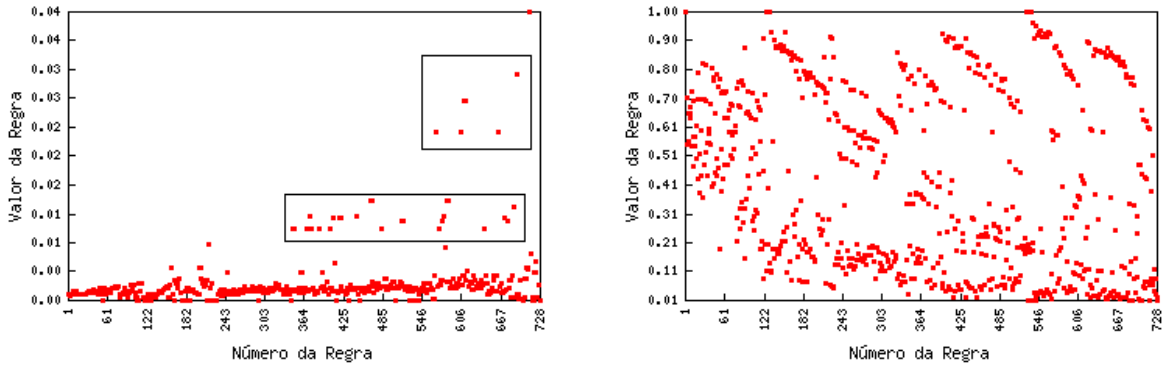


Figura 3.2: Processo de identificação de regras potencialmente interessantes no $\mathcal{ARInE}$ (Melanda, 2004)

Definido o conjunto foco, a próxima etapa consiste na aplicação de medidas objetivas para filtrar as regras e identificar aquelas potencialmente interessantes. Isso pode ser realizado utilizando gráficos de dispersão, de distribuição de freqüência ou de análise de

Pareto. No *ARInE* estão disponíveis diversas medidas objetivas que podem ser utilizadas nessa etapa. Para auxiliar o usuário na seleção das melhores medidas a serem utilizadas para cada conjunto de regras, o *ARInE* apóia a análise de similaridade das medidas, por meio da análise fatorial. Essa análise fatorial deve ser realizada com os valores das medidas no conjunto de regras que está sendo analisado, sendo possível verificar a similaridade entre as diferentes medidas e agrupar aquelas com comportamentos semelhantes. Com isso, pode-se analisar graficamente apenas uma medida de cada grupo para identificar subconjuntos de regras potencialmente interessantes.

A identificação desses pequenos subconjuntos pode ser realizada visualmente com a utilização de um gráfico de dispersão, como o apresentado na Figura 3.3(a), no qual podem ser verificados grupos de regras com valores mais altos para determinada medida. Contudo, nem sempre a identificação desses subconjuntos é trivial, como no caso apresentado na Figura 3.3(b), no qual os pontos estão dispersos. Nesse caso pode ser necessário o uso de outros métodos de análise.



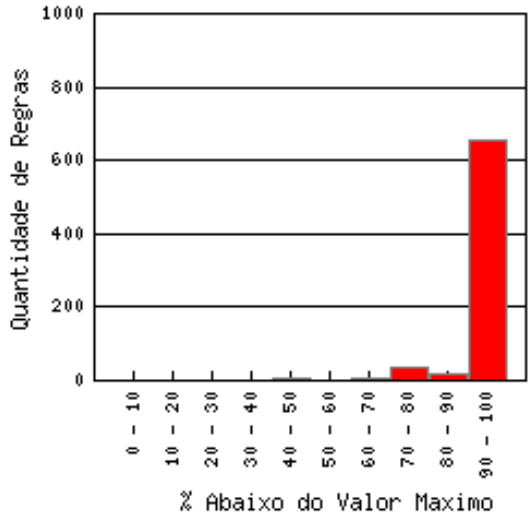
(a) Identificação trivial

(b) Identificação não trivial

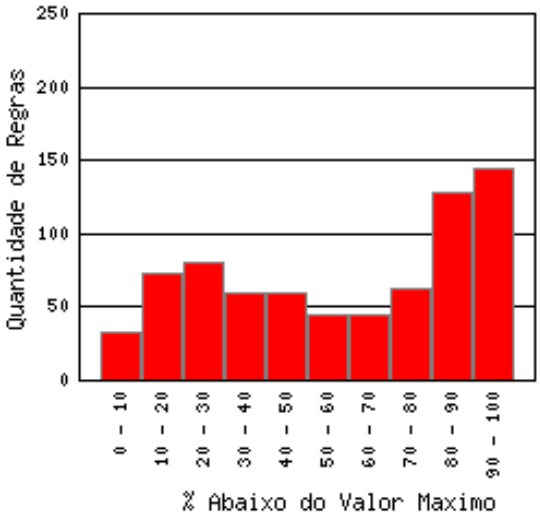
Figura 3.3: Gráficos de dispersão de valores de medidas para regras de associação

Na análise da distribuição de freqüência, os valores de uma dada medida são discretizados e é verificado o número de regras para cada intervalo. Na Figura 3.4 são exemplificados os resultados obtidos para esse tipo de análise. Esse procedimento visa analisar conjunta e comparativamente regras com valores semelhantes para determinada medida.

A análise de Pareto é baseada no Princípio de Pareto, que afirma: "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. No gráfico para análise de Pareto (Figura 3.5) é destacado como referência o percentual de 80% da soma dos valores da medida. Com base na curva do gráfico, o usuário pode definir o melhor percentual limite para corte e seleção das regras. No caso apresentado na Figura 3.5(a) a relação 80/20 não é válida, mas pode ser obtida a relação 60/30 que indica



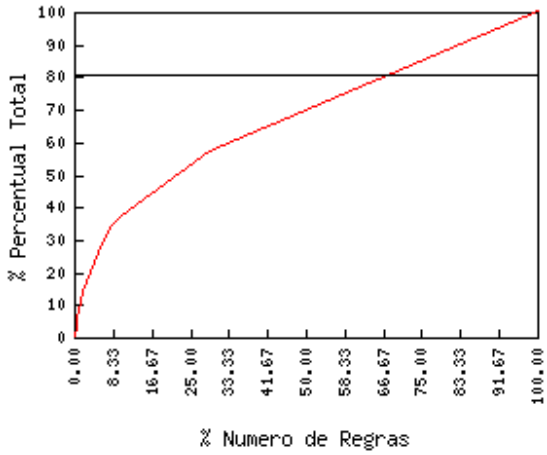
(a) Distribuição irregular de regras por valores de medidas



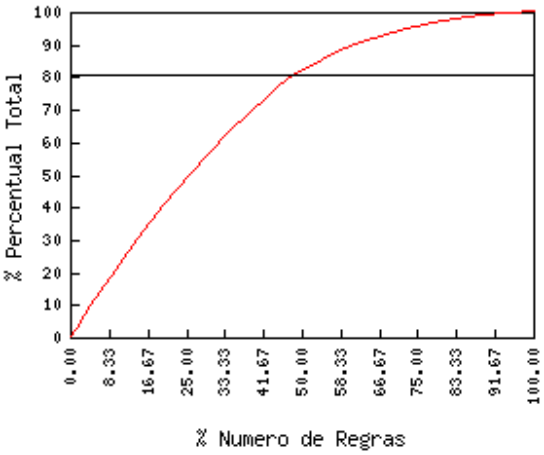
(b) Distribuição regular de regras por valores de medidas

Figura 3.4: Gráficos de distribuição de freqüência de valores de medidas para regras de associação

que para 60% de representatividade da medida são selecionadas apenas 30% das regras. Já na Figura 3.5(b) é obtida a relação 80/50 (para 80% de representatividade da medida, são selecionados 50% das regras).



(a) Curva com inclinação inicial mais acentuada



(b) Curva com inclinação inicial menos acentuada

Figura 3.5: Gráficos para análise de Pareto

Independente da análise gráfica utilizada (dispersão, distribuição de freqüência ou análise de Pareto), é possível reduzir ainda mais o número de regras a ser apresentado ao usuário especialista, por meio da intersecção dos subconjuntos de regras potencialmente interessantes identificados com o uso de diferentes medidas objetivas.

3.4.3 Aplicativo AIAS

Nessa seção é apresentado o aplicativo AIAS (*Association Interestingness Analysis System*), que utiliza medidas subjetivas para avaliar regras de associação. Esse aplicativo, desenvolvido por Liu, Hsu, Chen, & Ma (2000), permite identificar regras interessantes, em especial regras inesperadas, a partir de um conjunto de regras de associação. Para tanto, o AIAS emprega o conhecimento existente do domínio para ordenar as regras segundo diversos tipos de inesperabilidade.

A análise das regras no AIAS ocorre da seguinte maneira:

1. O usuário fornece o conhecimento do domínio.
2. O sistema analisa as regras descobertas verificando a conformidade e a inesperabilidade destas em relação ao conhecimento especificado pelo usuário.
3. O usuário utiliza os resultados das análises para visualizar as regras, salvar aquelas que são interessantes e remover as que não deseja.

O conhecimento que o usuário tem sobre o domínio pode ser fornecido ao sistema como sendo um conhecimento impreciso (*reasonably precise concept*) ou uma impressão geral (*general impression*). O conhecimento impreciso possibilita informar um conhecimento que o especialista supõe ser verdadeiro, já com uma impressão geral é possível informar uma relação que o especialista acredita existir entre os itens especificados. Esses dois tipos de conhecimentos são fornecidos pelo usuário utilizando uma linguagem definida para esse aplicativo, sendo expressos da seguinte maneira:

Conhecimento impreciso

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

Impressão geral

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

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 tanto na especificação de um conhecimento impreciso quanto na de uma impressão geral. 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.

Após o conhecimento ser especificado, o sistema o utiliza para analisar as regras descobertas. São realizadas análises sintáticas com cada conhecimento impreciso e impressão

geral fornecidos pelo usuário. 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, conforme foram apresentadas na Seção 3.3.2.

Calculados os valores das medidas, o usuário pode utilizá-los para visualizar as regras descobertas por diferentes aspectos (Liu, Hsu, Chen, & Ma, 2000):

Tipos de interesse As regras são categorizadas conforme 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).

Atributos interessantes A apresentação dos atributos interessantes pode ser considerada mais produtiva que a apresentação da regra como um todo.

3.5 Considerações Finais

A avaliação do conhecimento obtido na etapa de extração de padrões não é uma tarefa simples, visto que envolve um grande número de critérios, tanto objetivos quanto subjetivos. De maneira geral, o conhecimento deve ser avaliado quanto a seu desempenho, sua compreensibilidade e sua interessabilidade.

Nesse capítulo foi discutida a avaliação de regras, enfatizando a avaliação da interessabilidade. Foram apresentadas algumas técnicas para auxiliar a identificação de regras interessantes, entre elas a aplicação de medidas para avaliação de conhecimento. Essas medidas, objetivas ou subjetivas, são utilizadas em ferramentas também apresentadas nesse capítulo. O desenvolvimento de ferramentas como essas tem grande importância, especialmente no caso das regras de associação, pois o número de regras a serem avaliadas dificulta, ou até mesmo inviabiliza, uma avaliação manual.

O uso de medidas objetivas na identificação de conhecimento interessante é relevante, pois são medidas gerais e que podem ser aplicadas facilmente a qualquer conjunto de regras. Porém essas medidas podem não ser suficientes, já que não consideram o conhecimento, as intuições ou as expectativas do usuário no momento da avaliação. As medidas subjetivas consideram esses fatores, mas algumas limitações podem ser encontradas com o uso da subjetividade do usuário.

No próximo capítulo é proposta uma metodologia para identificação de regras de associação interessantes que combina o uso de medidas objetivas e subjetivas, aproveitando as vantagens que cada tipo oferece.

Capítulo

4

Metodologia para Identificação de Regras de Associação Interessantes

Como apresentado nos capítulos anteriores, a identificação de conhecimento interessante durante a etapa de pós-processamento é um dos fatores que determinam o sucesso do processo de mineração de dados. Porém, avaliar a interessabilidade do conhecimento descoberto é uma tarefa difícil, sendo, na maioria dos casos, muito subjetiva. Qualificar um conhecimento descoberto como interessante, ou não, depende, em parte, do conhecimento que já se tem sobre o domínio e dos objetivos que levaram à realização do processo de mineração de dados. No entanto, é necessário o desenvolvimento de métodos que auxiliem a identificação de conhecimento interessante da maneira mais geral possível, para que possam ser utilizados em diferentes domínios e que sejam adequados a diferentes objetivos e especialistas[1].

Nesse capítulo é apresentada uma metodologia para identificação de regras de associação interessantes, que utiliza duas das técnicas apresentadas na Seção 3.2: avaliação por consulta e aplicação de medidas de avaliação de conhecimento. O ponto chave dessa metodologia é o uso combinado de medidas objetivas e subjetivas, também discutido nesse capítulo. Além disso, é apresentado o módulo de exploração de regras de associação com medidas subjetivas (RulEE-SEAR), projetado e desenvolvido para viabilizar o uso da metodologia proposta no ambiente RulEE.

[1]Nesse capítulo, os termos "especialista do domínio" e "usuário" são utilizados de maneira indiscriminada, sendo que esses dois tipos de usuários do processo de mineração de dados podem realizar as atividades da metodologia proposta, visto qual é o mais indicado dados os objetivos do processo.

4.1 Combinação de Medidas Objetivas e Subjetivas

No Capítulo 3 foram apresentadas algumas técnicas que podem ser utilizadas para auxiliar a identificação de regras interessantes, entre elas está o uso de medidas de avaliação de conhecimento. De acordo com o que é considerado no cálculo dessas medidas, elas geralmente são classificadas como medidas objetivas ou medidas subjetivas. As medidas objetivas são calculadas com base nos dados e na estrutura das regras, enquanto o cálculo das medidas subjetivas considera, de alguma maneira, o conhecimento e/ou os interesses de um usuário.

Assim, as medidas objetivas têm a vantagem de serem mais gerais, podem ser utilizadas para avaliar regras de qualquer domínio e não dependem do especialista. Porém, apesar de eficientes em muitos casos, as medidas objetivas podem não levar à identificação de regras realmente interessantes ao usuário. Isso ocorre porque o grau de interesse que determinada regra possui pode variar dependendo do usuário que irá utilizá-la e de seu objetivo. Considere, por exemplo, um processo de mineração de dados realizado com uma base de dados de um supermercado. Em determinado período do ano o usuário pode estar interessado em regras que associem um produto x, pois seu objetivo é encontrar uma maneira vantajosa de organizar as prateleiras durante uma promoção desse produto. Em outro período do ano, o mesmo usuário pode estar interessado em regras que possuam um produto y, devido à um aumento do preço desse produto. Também deve-se notar que o conhecimento que um especialista tem sobre o domínio se altera constantemente, em função de seus estudos e experiências. Portanto, uma regra considerada interessante em uma determinada avaliação pode não ser interessante em outro momento, dado que o conhecimento prévio do especialista ou seus objetivos podem ter mudado. Essas variações podem ser captadas com o uso de medidas subjetivas, uma vez que essas medidas são calculadas com base no conhecimento e nas necessidades do usuário no momento da avaliação.

No entanto, para que as medidas subjetivas sejam utilizadas, é comum o usuário ter que fornecer seu conhecimento em uma linguagem pré-definida ou ter que quantificar conceitos imprecisos que possui, como indicar com valores numéricos o que significa para ele expressões do tipo "alto", "baixo", "grande" e "pequeno". Isso pode fazer com que o usuário tenha dificuldade em expressar seu conhecimento ao fornecê-lo a um sistema que o apóie no processo de identificação das regras interessantes.

Visando aproveitar as vantagens das medidas objetivas e das subjetivas, neste trabalho é proposta uma metodologia para identificação de regras de associação interessantes combinando o uso dessas medidas. Vale ressaltar que o foco deste trabalho são as medidas subjetivas e a sua combinação com as medidas objetivas, visto que a análise de regras de associação com medidas objetivas foram exploradas em um trabalho anterior (Melanda, 2004). Na próxima seção é apresentada a metodologia proposta neste trabalho.

4.2 Metodologia Proposta

A metodologia para identificação de regras de associação interessantes proposta neste trabalho tem como base as técnicas avaliação por consulta e aplicação de medidas de avaliação de conhecimento, descritas na Seção 3.2. A avaliação por consulta é utilizada no início do processo de identificação de regras interessantes com o objetivo de focar a análise em um conjunto de interesse do usuário. Logo após são utilizadas as medidas de avaliação, objetivas e subjetivas, de maneira a aproveitar as vantagens de cada uma durante o processo. Dessa maneira, primeiro são utilizadas as medidas objetivas, visando filtrar o conjunto selecionado, que geralmente é formado por um grande número de regras. Já as medidas subjetivas são utilizadas no fim do processo, auxiliando o usuário a analisar as regras de acordo com seus conhecimentos e objetivos.

A Figura 4.1 ilustra a metodologia proposta, que inicia a busca por regras interessantes com uma análise objetiva das regras de associação resultantes da extração de padrões. Essa análise é realizada com base em consultas e em medidas objetivas. Com as consultas é selecionado um conjunto de regras foco e com as medidas objetivas, um subconjunto deste com regras potencialmente interessantes. Esse subconjunto é utilizado na próxima etapa da metodologia, captação do conhecimento e dos interesses do especialista do domínio, na qual o especialista 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, que é 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 medidas subjetivas para guiá-lo na exploração das regras resultantes no conjunto foco e na identificação das regras interessantes. A seguir, são descritas em detalhes as quatro etapas da metodologia.

1. **Análise Objetiva**
   Essa etapa é realizada com o objetivo de filtrar o conjunto de regras resultante da extração de padrões. Busca-se identificar um pequeno subconjunto de regras para serem avaliadas por um especialista do domínio. Para tanto, é utilizado o fluxo de análise de regras de associação proposto por Melanda (2004) e utilizado no ambiente $\mathcal{ARInE}$, apresentado na Seção 3.4.2. Esse fluxo de análise é composto por consultas realizadas no conjunto de regras resultante da extração de padrões e aplicação de medidas objetivas. As consultas são realizadas caso o usuário deseje analisar determinado atributo ou item, formando um conjunto foco. Se o usuário não tem interesse em um item específico, o conjunto foco é formado por todo o

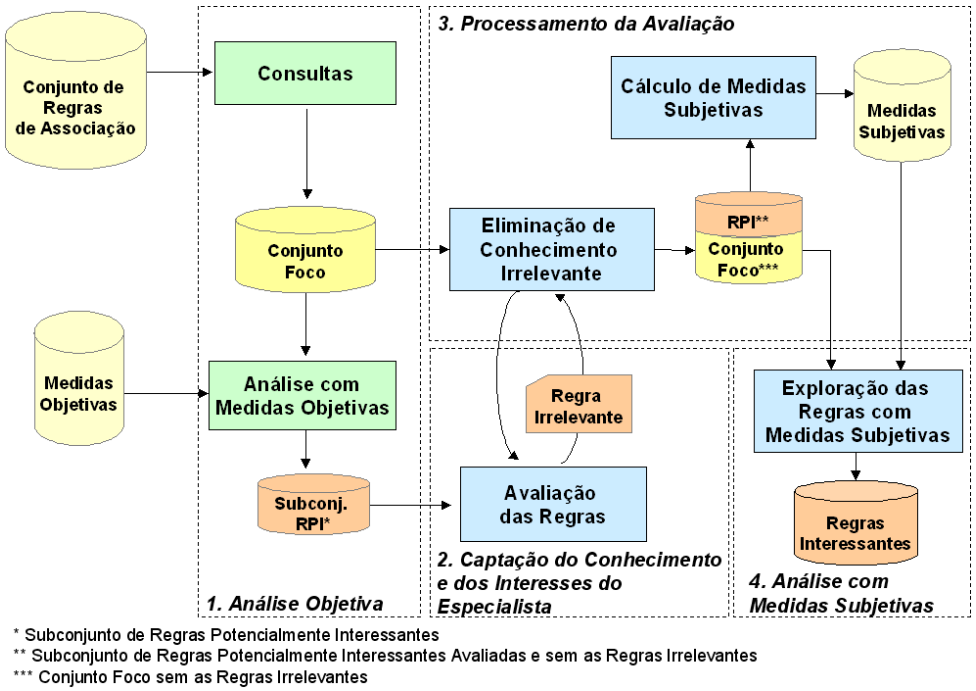


Figura 4.1: Metodologia para identificação de regras de associação interessantes

conjunto de regras. A partir da definição do conjunto foco inicia-se a análise com medidas objetivas. Para cada medida pode-se analisar graficamente a distribuição de seus valores para o universo de regras do conjunto foco de três maneiras: gráfico de dispersão, gráfico de distribuição de freqüência e gráfico de análise de Pareto. Analisando as medidas objetivas define-se um valor de corte para filtrar as regras do conjunto foco e obter, para cada uma dessas medidas, um subconjunto de regras. A intersecção desses subconjuntos forma o subconjunto de regras potencialmente interessantes (RPI) e é apresentado ao especialista do domínio, dando início a etapa de captação do conhecimento e dos interesses do especialista, com a avaliação dessas regras.

2. **Captação do Conhecimento e dos Interesses do Especialista**
   Essa etapa pode ser vista como uma entrevista com o especialista do domínio, na qual ele responde um questionário simples, avaliando as regras do subconjunto RPI. Para otimizar a avaliação, as regras são ordenadas pelo tamanho do *itemset* que as constituem e são apresentadas em ordem crescente. O especialista deve indicar uma ou mais opções de avaliação para cada regra do subconjunto RPI. Cada opção de avaliação classifica o conhecimento representado pela regra, sendo elas: conhecimento inesperado, conhecimento útil, conhecimento óbvio do domínio, conhecimento prévio do especialista e conhecimento irrelevante de acordo com os objetivos da análise. Essas classificações que o especialista pode atribuir a cada regra ao avaliá-la são definidas como se segue:

   Conhecimento inesperado Indica aquele conhecimento que é uma novidade para o especialista, algo em que ele não havia pensado ou que contradiz um conhe-

cimento prévio que ele tenha.

Exemplo: Considerando uma base de dados de compras realizadas em um supermercado, a regra *adoçante* ⇒ *leite_condensado* pode ser considerada inesperada por determinado especialista que acreditava que raramente esses dois produtos eram vendidos juntos.

Conhecimento útil Indica aquele conhecimento que pode ser utilizado para tomar alguma decisão.

Exemplo: Ainda no domínio de compras em um supermercado e considerando que o objetivo do especialista é reorganizar suas prateleiras durante uma promoção do produto *adoçante*. A regra *adoçante* ⇒ *leite_condensado* também poderia ser considerada útil pelo especialista, pois com esse conhecimento ele poderia colocar o produto *leite_condensado* próximo ao produto *adoçante* visando aumentar a venda de *leite_condensado* durante essa promoção.

Conhecimento prévio do especialista Indica o conhecimento formado de acordo com as experiências do especialista, podendo ser diferente para cada especialista.

Exemplo: A regra *leite_pó* ⇒ *vinho* pode ser considerada um conhecimento prévio para determinado especialista devido ao seu conhecimento sobre seus clientes ou devido à um processo de mineração de dados anterior que tenha encontrado esse padrão em seus dados.

Conhecimento óbvio do domínio Indica o conhecimento bem consolidado no domínio. Assim, todo conhecimento óbvio do domínio é também um conhecimento prévio do especialista, porém o conhecimento óbvio, em geral, não é diferente para cada especialista. Vale dizer que esse conhecimento nem sempre é óbvio para pessoas não especialistas no domínio.

Exemplo: A regra *shampoo* ⇒ *condicionador* pode ser considerada um conhecimento óbvio, visto que os comerciantes sabem que o produto *condicionador* freqüentemente é vendido junto com o produto *shampoo*.

Conhecimento irrelevante Indica aquele conhecimento que não é importante ou não é necessário de acordo com os objetivos definidos para o processo de mineração de dados durante a fase de identificação do problema.

Exemplo: Considerando o domínio de compras realizadas em um supermercado e o objetivo de encontrar os produtos que com freqüência são vendidos junto com o produto *shampoo* para, por exemplo, alterar a disposição desses produtos no supermercado e aumentar a venda conjunta dos mesmos. A regra *shampoo* ⇒ *condicionador* também pode ser considerada irrelevante pelo especialista, dado que, no supermercado em questão, esses dois produtos se encontram na mesma prateleira.

Para as regras classificadas como conhecimento irrelevante, o usuário deve fornecer

os itens que as tornam irrelevantes, para que seja realizada a eliminação dessas regras e de outras também irrelevantes, como parte da etapa de processamento da avaliação. Essa eliminação é realizada para que não seja apresentada outra regra irrelevante pelos mesmos motivos já avaliados. Com o fim da avaliação é iniciado o cálculo das medidas subjetivas, também na etapa de processamento da avaliação.

3. **Processamento da Avaliação**
   Dadas as respostas do usuário para cada regra pertencente ao subconjunto RPI é iniciado o processamento da avaliação nas regras do conjunto foco, definido durante a análise objetiva. Esse processamento consiste na eliminação das regras que representam conhecimento irrelevante e no cálculo de medidas subjetivas, sendo realizado da seguinte maneira:

   - A cada regra classificada como conhecimento irrelevante pelo especialista são excluídas do conjunto foco todas as regras que representam um conhecimento irrelevante de acordo com os itens indicados pelo usuário durante a avaliação. Primeiramente são eliminadas as regras que possuem os itens indicados na mesma posição em que aparecem na regra avaliada. Por exemplo, se o usuário avaliar a regra $a, b \Rightarrow c, d$ como irrelevante e indicar os itens a do antecedente e d do conseqüente, são excluídas todas as regras do conjunto foco que apresentam a no antecedente e d no conseqüente. Após essa primeira eliminação é verificado se existe alguma regra no conjunto foco com esses itens em outras posições. Em caso positivo, o usuário é questionado quanto à irrelevância dessas regras. Como citado anteriormente, esse processo de eliminação é realizado para que nenhuma outra regra que represente conhecimento irrelevante por um motivo já avaliado seja apresentada ao usuário.
   - Para as regras que não foram eliminadas do conjunto foco e que não pertencem ao subconjunto RPI são calculadas as medidas subjetivas definidas por Liu et al. (2000), que são: conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados. Esse cálculo é realizado utilizando cada regra classificada como conhecimento óbvio, prévio, inesperado ou útil durante a etapa anterior como um conhecimento impreciso[2] fornecido pelo usuário. As fórmulas utilizadas no cálculo dessas medidas foram apresentadas na Seção 3.3.2. Detalhes sobre o cálculo nessa metodologia podem ser vistos na Seção 4.4.4.

   Após o processamento, o especialista pode iniciar a análise das regras do conjunto foco resultante utilizando as medidas subjetivas.

4. **Análise com Medidas Subjetivas**
   Nessa etapa o especialista pode explorar as regras do conjunto foco resultante uti-

[2] A definição de conhecimento impreciso encontra-se na Seção 3.4.3.

lizando as medidas subjetivas como um guia, acessando as regras de acordo com cada medida e em relação a cada regra avaliada. Essa exploração deve ser conduzida de acordo com o objetivo do usuário no momento da avaliação. Por exemplo, se o usuário quer confirmar o seu conhecimento prévio, ele pode utilizar a medida conformidade, e listar as regras em conformidade com aquelas que ele avaliou como conhecimento óbvio ou prévio; já se seu objetivo é identificar conhecimento novo, ele pode usar as medidas antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados, e listar as regras inesperadas em relação a uma regra já avaliada. Ao navegar pelas regras, o usuário pode salvar aquelas que achar interessante. Assim, no final da análise, ele terá um conjunto contendo somente as regras que foram avaliadas como interessantes.

4.3 Considerações sobre a Metodologia

Conforme apresentado na seção anterior, a metodologia proposta combina o uso de medidas objetivas e subjetivas para auxiliar a identificação de regras de associação interessantes. A utilidade dessa combinação é um dos aspectos que devem ser ressaltados sobre essa metodologia, sendo os outros a abrangência da metodologia e a maneira como a subjetividade do usuário é captada. Esses três aspectos são discutidos a seguir:

- O uso combinado de medidas objetivas e subjetivas, abordado na Seção 4.1, pode ser de grande auxílio na identificação de regras de associação interessantes visto que com as medidas objetivas pode-se filtrar as regras, diminuindo o número de regras a serem avaliadas pelo especialista e as medidas subjetivas podem ser utilizadas para ordenar as regras de acordo com o interesse do usuário, facilitando a identificação de regras interessantes.

- A metodologia proposta utiliza o fluxo de análise objetiva proposto por Melanda (2004) combinado com as medidas subjetivas propostas por Liu, Hsu, Chen, & Ma (2000). Com isso, ela abrange dois fatores que tornam uma regra de associação interessante: a presença de itens interessantes e a presença de associações interessantes entre os itens. As regras que apresentam itens que interessam aos usuários podem ser selecionadas por meio das consultas, na análise objetiva, enquanto as regras com associações interessantes podem ser identificadas com o uso das medidas subjetivas.

- Como a subjetividade do usuário (seus conhecimentos e interesses) é captada por meio de um questionário, com o qual o usuário avalia algumas regras, algumas das limitações constantemente presentes em avaliações subjetivas podem ser suprimidas. Esse é o caso das limitações de linguagem e incerteza nos critérios (apresentadas na Seção 3.1.2), que não aparecem com o uso dessa metodologia, já que o usuário não

tem que fornecer explicitamente seu conhecimento para que este seja utilizado no cálculo das medidas subjetivas.

Outro ponto que deve ser ressaltado na metodologia proposta é a importância da identificação de conhecimento irrelevante durante a etapa de captação do conhecimento e dos interesses do especialista, dado que quando uma regra é avaliada como conhecimento irrelevante outras regras também são eliminadas do conjunto foco. Essa eliminação de regras irrelevantes auxilia a identificação de regras interessantes durante a análise com medidas subjetivas, visto que diminui a quantidade de regras a serem analisadas. No entanto, caso todas as regras do subconjunto RPI forem consideradas irrelevantes não será possível o cálculo das medidas subjetivas e deve ser realizada uma nova análise objetiva para formar um novo subconjunto RPI a ser avaliado.

A metodologia proposta pode ser utilizada no ambiente RulEE, por meio de seus módulos de pós-processamento $\mathcal{ARInE}$, apresentado na Seção 3.4.2, e RulEE-SEAR, descrito na próxima seção.

4.4 *Módulo de Exploração de Regras de Associação com Medidas Subjetivas* - RulEE-SEAR

O módulo de exploração de regras de associação com medidas subjetivas, RulEE-SEAR, foi desenvolvido como um dos módulos de pós-processamento do ambiente RulEE, com o objetivo de viabilizar a aplicação da metodologia apresentada nesse capítulo. Assim, essa metodologia pode ser utilizada por meio de dois módulos de pós-processamento do RulEE: o $\mathcal{ARInE}$, descrito na Seção 3.4.2, e o RulEE-SEAR. No $\mathcal{ARInE}$ é realizada a análise objetiva, primeira etapa da metodologia. Já no RulEE-SEAR, são realizadas as demais etapas (etapas 2, 3 e 4 da Figura 4.1), que compõem o que é chamado nesse módulo de processo de exploração subjetiva. Como a participação dos usuários nesse processo de exploração subjetiva é de extrema importância, o RulEE-SEAR possibilita a tradução de regras para uma linguagem mais fácil de ser compreendida pelos usuários.

Nessa seção é apresentado o RulEE-SEAR, descrevendo sua implementação e sua utilização. Vale ressaltar que, por ser um módulo do RulEE, o RulEE-SEAR utiliza seus repositórios e métodos para acesso aos dados. A seguir são apresentados os documentos elaborados durante a fase de análise e projeto do desenvolvimento do módulo, bem como as alterações realizadas nos repositórios e nas classes do RulEE, necessárias para a implementação do RulEE-SEAR. Também são apresentados nessa seção os procedimentos adotados para o cálculo das medidas subjetivas e para a tradução das regras. Por fim, é descrito como um processo de exploração subjetiva pode ser realizado no RulEE-SEAR.

4.4.1 Requisitos e Casos de Uso

A primeira atividade realizada para o desenvolvimento do RulEE-SEAR foi o levantamento de requisitos. Com esse levantamento definiu-se as funcionalidades desse módulo, que são:

- Criação, manipulação e finalização de um processo de exploração subjetiva.
- Avaliação de regras potencialmente interessantes, para a captação do conhecimento e dos interesses do especialista.
- Eliminação de regras consideradas irrelevantes.
- Cálculo das medidas subjetivas conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados.
- Análise das regras de um conjunto foco com medidas subjetivas.
- Visualização de explorações realizadas, para apoiar o compartilhamento de conhecimento entre usuários de um mesmo projeto.
- Definição de traduções para um conjunto de regras, visando facilitar a compreensão das regras durante o processo de exploração subjetiva.

A partir da realização dessa atividade foi elaborado um documento de requisitos, que é apresentado no Apêndice A. Após a elaboração desse documento foram definidos os casos de uso do RulEE-SEAR, apresentados na Figura 4.2. Detalhes sobre esses casos de uso podem ser encontrados nas especificações apresentadas no Apêndice B.

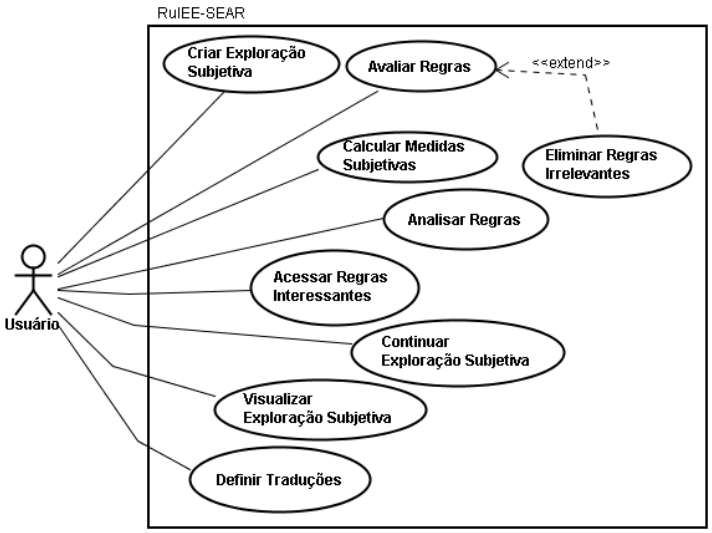


Figura 4.2: Casos de uso do RulEE-SEAR

4.4.2 Repositório

Para a implementação do módulo RulEE-SEAR foi necessária a criação de seis novas tabelas nos repositórios do RulEE: SUBJECTIVE_EXPLORATION, SUB_FOCUS_RPI, SUB_EVALUATION, SUB_MEASURE_VALUE, SUB_TRANSLATION_RULE e SUB_-TRANSLATION_ITEM. Uma breve descrição de cada uma dessas tabelas encontra-se na Tabela 4.1 e na Figura 4.3 é apresentada parte do diagrama físico da base de dados do ambiente RulEE, contendo as tabelas específicas do RulEE-SEAR e as principais tabelas do ambiente RulEE utilizadas por esse módulo.

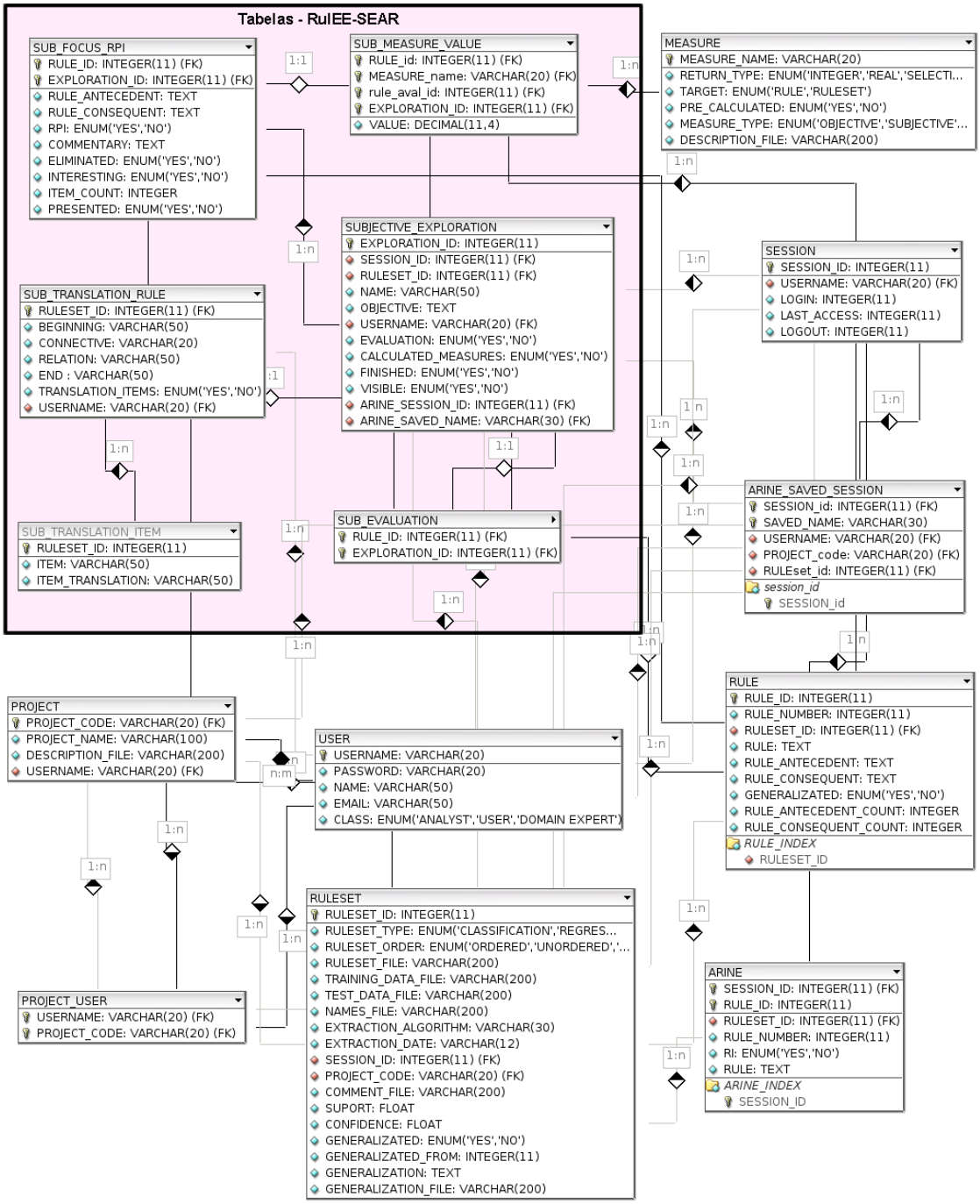


Figura 4.3: Parte do diagrama físico da base de dados do RulEE

A descrição completa das tabelas criadas e de seus atributos encontra-se em um relatório técnico sobre o RulEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006).

Nome da Tabela	Descrição
SUBJECTIVE_EXPLORATION	Armazena dados referentes a um processo de exploração subjetiva iniciado no RulEE-SEAR, como nome e objetivo da exploração, usuário responsável, código do conjunto de regras utilizado e etapas já realizadas.
SUB_FOCUS_RPI	Armazena as regras do conjunto foco e, conseqüentemente, do subconjunto de regras potencialmente interessantes formados em uma análise no $\mathcal{ARInE}$. Nessa tabela também são armazenados os comentários sobre as regras escritos pelo usuário e a informação se a regra foi considerada interessante ou não.
SUB_EVALUATION	Armazena as avaliações das regras do subconjunto RPI realizadas pelo usuário
SUB_MEASURE_VALUE	Armazena os valores das medidas subjetivas que foram calculadas.
SUB_TRANSLATION_RULE	Armazena os textos para tradução de uma regra de determinado conjunto de regras, sendo eles: texto inicial, conectivo de itens, texto entre antecedente e conseqüente e texto final.
SUB_TRANSLATION_ITEM	Armazena as traduções para os itens presentes em regras de um conjunto de regras.

Tabela 4.1: Descrição das tabelas criadas para o RulEE-SEAR

4.4.3 Classes

Para a implementação do RulEE-SEAR também foi necessária a criação de um novo método na classe Project do RulEE. Essa classe representa um projeto existente no ambiente e é utilizada para retornar as informações sobre esse projeto. O método criado, GetRulesetsProject, retorna os conjuntos de regras inseridos em um determinado projeto.

Além disso, foram criadas as classes SEAR e Translation. A classe SEAR contém todos os métodos utilizados para manipular um processo de exploração subjetiva. Os principais métodos dessa classe e suas funcionalidades são apresentados na Tabela 4.2. Já a classe Translation contém os métodos para inserir e recuperar traduções de determinado conjunto de regras. Detalhes sobre a alteração realizada na classe Project e sobre as classes SEAR e Translation podem ser encontradas no relatório técnico em que é descrito o RulEE-SEAR (Sinoara, Fujimoto, & Rezende, 2006).

4.4.4 Cálculo das Medidas Subjetivas

As medidas subjetivas utilizadas na metodologia proposta e, conseqüentemente, no RulEE-SEAR são conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados. As equações utilizadas para calcular essas medidas

Nome do Método	Funcionalidade
SaveExploration	Salva os dados de uma exploração na base de dados, ou seja, cria um processo de exploração subjetiva.
GetNextRule	Encontra a próxima regra a ser avaliada pelo usuário.
SaveEvaluation	Salva as informações fornecidas pelo usuário ao avaliar uma regra.
EliminateIrrelevantRules	Elimina as regras consideradas irrelevantes pelo usuário.
CheckPossibleIrrelevantRules	Verifica se exite alguma outra regra apresentando os itens indicados como irrelevantes em posições diferentes em relação à regra avaliada.
SetPresentedRules	Marca as regras que já foram apresentadas ao usuário, durante o processo de eliminação de regras irrelevantes e que não foram consideradas irrelevantes.
CalculateSubjectiveMeasures	Calcula os valores das medidas subjetivas.
GetVisibleExplorations	Seleciona as explorações que podem ser visualizadas por determinado usuário em um determinado projeto. Isso inclui todas as explorações do usuário que estejam encerradas, além das explorações visíveis e encerradas de outros usuários.
GetRulesByMeasure	Dada uma regra avaliada, uma medida e um intervalo de valores, retorna as regras que em relação a essa regra avaliada apresenta o valor da medida no intervalo definido.

Tabela 4.2: Descrição dos principais métodos da classe SEAR

são apresentadas na Seção 3.3.2 (equações 3.13, 3.14, 3.15 e 3.16, respectivamente), sendo que os valores de L_{ij} e R_{ij} são obtidos de maneira distinta de acordo com o tipo de conhecimento fornecido pelo usuário (impressão geral ou conhecimento impreciso). Na metodologia proposta, as regras avaliadas pelo usuário na etapa de captação do conhecimento e dos interesses do especialista são utilizadas no cálculo das medidas subjetivas como um conhecimento impreciso. Dado um conhecimento impreciso, os valores de L_{ij} e R_{ij} são obtidos pelas equações 4.1 e 4.2[3]:

$$L_{ij} = min\left(\frac{LM_{ij}}{LN_i}, \frac{LSM_{ij}}{LSN_j}\right) \tag{4.1}$$

$$R_{ij} = min\left(\frac{RM_{ij}}{RN_i}, \frac{RVM_{ij}}{RVN_j}\right) \tag{4.2}$$

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 $\frac{LSM_{ij}}{LSN_j} = 1$

[3]As equações 4.1 e 4.2 também são apresentadas na Seção 3.3.2 (equações 3.21 e 3.22) e foram reapresentadas nessa seção para melhor compreensão de seu conteúdo.

(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.

Como apresentado na Seção 3.4.3, o conhecimento do usuário pode ser especificado utilizando-se classes de uma taxonomia. Assim, para o cálculo dos valores de L_{ij} e R_{ij} deve-se considerar que um item a de uma regra descoberta casa com um elemento S definido pelo usuário quando:

- S é um item existente na base de dados e $a = S$ ou;

- S é uma classe C e a é uma instância de C ou;

- S é da forma $C+$ ou $C*$, e $a \in C$.

No RuLEE-SEAR, as equações para o cálculo dos valores de L_{ij} e R_{ij} foram adaptadas, visto que regras generalizadas e taxonomias não são foco desse módulo. Assim, os valores de LM_{ij} e RM_{ij} foram considerados iguais aos valores de LSM_{ij} e RVM_{ij}, respectivamente. Portanto, para uma regra A_j avaliada pelo usuário e uma regra F_i do conjunto foco que não tenha sido eliminada e que não pertença ao subconjunto RPI, os valores de L_{ij} e R_{ij} são dados pelas equações 4.3 e 4.4:

$$L_{ij} = min\left(\frac{AntFA_{ij}}{AntF_i}, \frac{AntFA_{ij}}{AntA_j}\right) \tag{4.3}$$

$$R_{ij} = min\left(\frac{ConsFA_{ij}}{ConsF_i}, \frac{ConsFA_{ij}}{ConsA_j}\right) \tag{4.4}$$

sendo que,

- $AntFA_{ij}$ e $ConsFA_{ij}$ representam, cada qual, o número de itens no antecedente e no conseqüente de A_j que casam, respectivamente, com o antecedente e o conseqüente de F_i;

- $AntF_i$ e $ConsF_i$ representam, respectivamente, o número de itens no antecedente e no conseqüente de F_i;

- $AntA_j$ e $ConsA_j$ representam, respectivamente, o número de itens no antecedente e no conseqüente de A_j.

Assim, para cada regra do conjunto foco, desconsiderando as regras eliminadas e as regras do subconjunto RPI, são calculados os valores das quatro medidas em relação a cada regra avaliada (exceto as regras avaliadas como conhecimento irrelevante), como apresentado no Algoritmo 1 (no qual, a notação ||X|| representa o número de elementos de X).

4.4.5 Tradução de Regras

Uma funcionalidade importante para facilitar a participação do usuário no processo de exploração subjetiva realizada no RulEE-SEAR é a tradução de regras. Conforme apresentado na Seção 2.2, as regras de associação são facilmente compreendidas, porém a compreensibilidade de uma regra não depende apenas de sua estrutura, mas também da semântica da regra e dos itens presentes. Por exemplo, o especialista do domínio pode não conhecer a nomenclatura dos atributos (ou itens) utilizada na base de dados, prejudicando seu entendimento das regras.

Assim, para cada conjunto de regras de associação presente no RulEE, é possível inserir no RulEE-SEAR dois tipos de traduções: tradução da regra e tradução dos itens. A tradução da regra diz como as regras desse conjunto devem ser interpretadas. As informações necessárias para traduzir as regras de determinado conjunto são: início (texto que deve aparecer antes do antecedente das regras), conectivo (texto de ligação entre itens do antecedente ou do conseqüente), relação (texto que indica a relação entre o antecedente e o conseqüente) e fim (texto que deve aparecer depois do conseqüente das regras). Com essas informações fornecidas para um conjunto de regras, todas as regras desse conjunto podem ser traduzidas como apresentado no Exemplo 1.

Exemplo 1

Regra 1: $pao \ \& \ leite \Rightarrow margarina$
Regra 1 traduzida: Quem compra pao e leite também compra margarina.

Regra 2: $shampoo \Rightarrow condicionador$
Regra 2 traduzida: Quem compra shampoo também compra condicionador.

Informações fornecidas para tradução:

- Início: Quem compra
- Conectivo: e
- Relação: também compra
- Fim: .

As informações necessárias para traduzir uma regra podem ser obtidas a partir do objetivo do processo de mineração de dados definido na fase de identificação do problema.

Algoritmo 1 Cálculo das medidas subjetivas no RulEE-SEAR

Requisito: Conjunto A de regras avaliadas pelo usuário, exceto as regras avaliadas como conhecimento irrelevante e conjunto F de regras do conjunto foco não eliminadas e não pertencentes ao subconjunto RPI.
Assegura: Valores das medidas subjetivas para cada regra de F em relação a cada regra de A.

```
para todo regra A_j ∈ A faça
  antA ← antecedentes de A_j {Cria um vetor com os itens do antecedente da regraA_j}
  consA := conseqüentes de A_j {Cria um vetor com os itens do conseqüente da regra A_j}
  para todo regra F_i ∈ F faça
    antF ← antecedentes de F_i {Cria um vetor com os itens do antecedente da regra F_i}
    consF ← conseqüentes de F_i {Cria um vetor com os itens do conseqüente da regra F_i}

    {Verifica quantos itens do antecedente de A_j casam com os itens do antecedente de F_i}
    antFA ← 0 {Número de itens casados nos antecedentes}
    para todo item itemA ∈ antA faça
      k ← 0
      enquanto antF[k] ≠ itemA e k < ||antF|| faça
        k ← k + 1
      fim-enquanto
      se antF[k] = itemA então
        antFA ← antFA + 1
      fim-se
    fim-para

    {Verifica quantos itens do conseqüente de A_j casam com os itens do conseqüente de F_i}
    consFA ← 0 {Número de itens casados nos conseqüentes}
    para todo item itemA ∈ consA faça
      k ← 0
      enquanto consF[k] ≠ itemA e k < ||consF|| faça
        k ← k + 1
      fim-enquanto
      se consF[k] = itemA então
        consFA ← antFA + 1
      fim-se
    fim-para

    {Calcula os valores de L_ij e R_ij}
    se ||antA|| > ||antF|| então
      L_ij ← antFA/||antA||
    senão
      L_ij ← antFA/||antF||
    fim-se
    se ||consA|| > ||consF|| então
      R_ij ← consFA/||consA||
    senão
      R_ij ← consFA/||consF||
    fim-se

    {Calcula os valores das medidas subjetivas}
    confm_ij ← L_ij * R_ij {conformidade}
    aux ← R_ij − L_ij {antecedente inesperado}
    se aux > 0 então
      unexpCond_ij ← aux
    senão
      unexpCond_ij ← 0
    fim-se
    aux ← L_ij − R_ij {conseqüente inesperado}
    se aux > 0 então
      unexpConseq_ij ← aux
    senão
      unexpConseq_ij ← 0
    fim-se
    bsUnexp_ij ← 1 − maior(confm_ij, unexpCond_ij, unexpConseq_ij) {antecedente e conseqüente inesperados}
  fim-para
fim-para
```

Nessa fase, ao ser selecionada a tarefa associação também foi definido o tipo de conhecimento que se pretendia obter com o processo de mineração de dados e, conseqüentemente, a semântica de uma regra de associação que seria gerada.

Já a tradução dos itens visa esclarecer ao especialista o significado de cada item. Essa

tradução é opcional. No caso apresentado no Exemplo 1 não foi necessária a tradução dos itens, pois qualquer especialista que fosse analisar as regras entenderia o significado de *pao*, *leite*, *margarina*, *shampoo* e *condicionador*. No entanto, nos casos em que o especialista não tem conhecimento da nomenclatura dos itens presentes na base de dados é necessária a tradução dos mesmos, como no Exemplo 2. A informação necessária para traduzir os itens também pode ser obtida durante a identificação do problema, a partir do conhecimento adquirido sobre a base de dados.

Exemplo 2

Regra: $temperatura = [27 - 30]$ & $umidade = alta \Rightarrow chuva = presente$

Regra traduzida: A ocorrência de temperatura alta e umidade alta leva a ocorrência de chuva.

Informações fornecidas para tradução:

- Início: A ocorrência de
- Conectivo: e
- Relação: leva a ocorrência de
- Fim: .

- Tradução do item $temperatura = [27 - 30]$: temperatura alta
- Tradução do item $umidade = alta$: umidade alta
- Tradução do item $chuva = presente$: chuva

4.4.6 Descrição do Processo de Exploração Subjetiva no RulEE-SEAR

Nessa seção é descrita a realização de um processo de exploração subjetiva no RulEE-SEAR, sendo apresentadas as principais telas de sua interface. Como já mencionado, o processo de exploração subjetiva compreende as etapas 2, 3 e 4 da metodologia proposta. Assim, antes de iniciar a exploração no RulEE-SEAR, a primeira etapa já deve ter sido realizada no $\mathcal{ARInE}$, no qual o usuário deve salvar um conjunto foco, um subconjunto de regras potencialmente interessantes, bem como a sessão de análise.

O acesso ao RulEE-SEAR é realizado por meio da interface do ambiente RulEE (http://143.107.231.137/rulee/index.html) e assim são mantidas as vantagens adquiridas com a utilização da Web, que são a independência de plataforma e de localização física de seus usuários. Portanto, é possível que o analista do processo de mineração de dados e o especialista do domínio estejam em lugares distintos durante a execução do processo.

Antes do RulEE-SEAR ser acessado deve ser selecionado o projeto com o qual se quer trabalhar. Com esse projeto selecionado, o usuário pode iniciar um novo processo de exploração, dar continuidade a um processo que ele tenha criado anteriormente, visualizar

uma exploração já encerrada ou inserir tradução para um conjunto de regras. Selecionando a opção de iniciar um novo processo, o usuário deve fornecer os dados necessários para a criação desse processo, como ilustrado na Figura 4.4. A sessão salva no $\mathcal{ARInE}$ indica qual conjunto foco e subconjunto de regras potencialmente interessantes devem ser utilizados. O usuário também deve informar se esse processo de exploração poderá ser visualizado por outros usuários do mesmo projeto. Caso a exploração fique disponível a outros usuários, as informações fornecidas no nome e no objetivo da exploração são importantes para que em futuras visualizações os usuários saibam em qual contexto esse processo foi realizado.

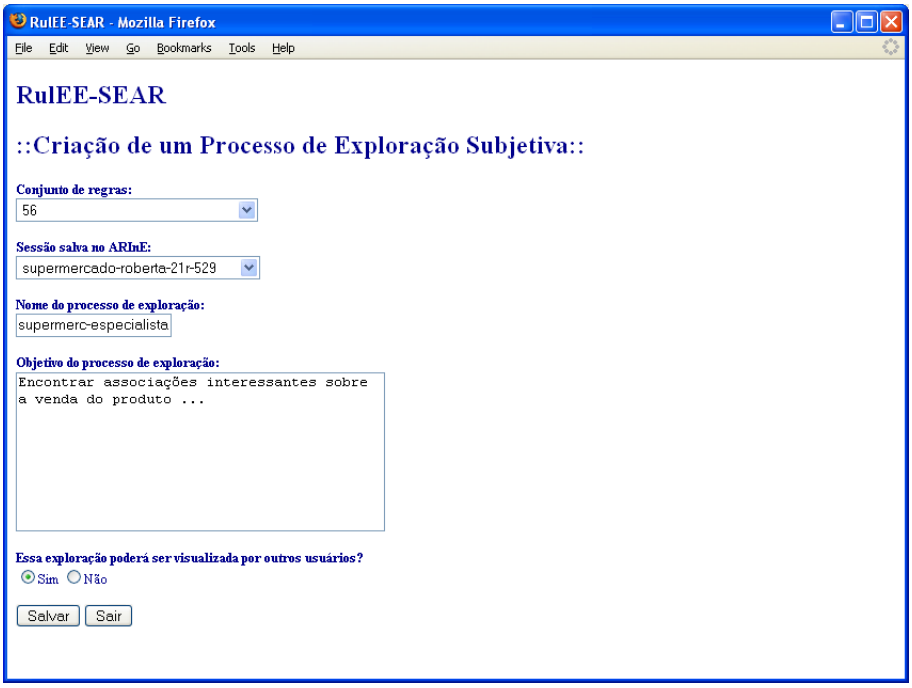


Figura 4.4: Criação de um processo de exploração no RulEE-SEAR

Com um processo de exploração criado, o próximo passo é a avaliação das regras do subconjunto RPI, ilustrado na Figura 4.5. Nessa etapa, as regras a serem avaliadas são ordenadas pelo tamanho do *itemset*, para que as regras menores sejam apresentadas no início da avaliação. Essa ordem facilita a avaliação para o usuário, pois, em geral, quanto menor a regra mais simples ela é. Assim, as regras são apresentadas uma a uma ao usuário, que pode classificá-las como conhecimento inesperado, útil, óbvio do domínio, prévio do especialista ou irrelevante, marcando uma ou mais dessas opções.

Ao avaliar uma regra como irrelevante, o usuário deve indicar os itens que tornam essa regra irrelevante (Figura 4.6). Caso a regra seja formada por apenas dois itens o RulEE-SEAR não questiona o usuário e considera os dois itens irrelevantes. Dados os itens irrelevantes, essa regra e as demais regras do conjunto foco (inclusive as regras do subconjunto RPI) que apresentam tais itens na mesma posição (antecedente ou conseqüente) da regra avaliada são eliminadas. As regras eliminadas não são mais apresentadas ao usuário nas demais etapas do processo de exploração subjetiva. Realizada essa eliminação, o RulEE-SEAR verifica se há regras no conjunto foco que apresenta esses itens em posi-

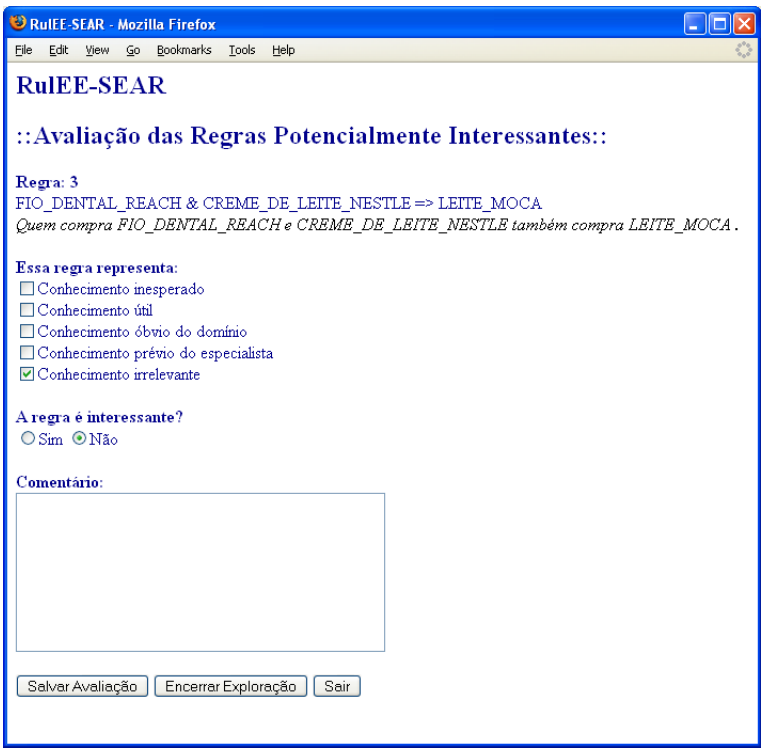


Figura 4.5: Avaliação das regras no RulEE-SEAR

ções diferentes da regra avaliada. Em caso afirmativo, as regras encontradas são ordenadas pelo tamanho do *itemset* e a menor delas é apresentada ao usuário (Figura 4.7), que deve informar se essa nova regra também é irrelevante. Caso o usuário a considere irrelevante, todo o processo de eliminação é repetido para essa regra. Caso contrário, a regra e as demais que apresentem os itens nas mesmas posições são marcadas como já apresentadas ao usuário. Enquanto existir outra regra com os itens indicados que não tenha sido eliminada ou marcada como já apresentada ao usuário, a regra é apresentada para seja informado se ela é irrelevante ou não e o processamento é realizado novamente. Quando nenhuma outra regra for encontrada, uma nova regra do subconjunto RPI é apresentada ao usuário para avaliação, como ilustrado na Figura 4.5.

Após a avaliação de todas as regras potencialmente interessantes, o usuário pode solicitar o cálculo das medidas subjetivas, que é realizado conforme o procedimento descrito na Seção 4.4.4.

Com as medidas subjetivas calculadas, pode ser iniciada a análise com essas medidas. Primeiramente são apresentadas as regras do subconjunto RPI divididas em dois grupos, regras avaliadas como conhecimento inesperado ou útil e regras avaliadas como conhecimento óbvio ou prévio, como ilustrado na Figura 4.8. Com as informações das avaliações de cada regra, o usuário pode selecionar uma das regras para, a partir dela, ter acesso às regras do conjunto foco por meio das medidas subjetivas. Após selecionar uma das regras, o usuário deve indicar qual medida quer utilizar e qual intervalo de valores lhe interessa, como ilustrado na Figura 4.9. Fornecidos esses valores, o RulEE-SEAR apresenta as regras do conjunto foco que satisfazem a seguinte condição: em relação à regra

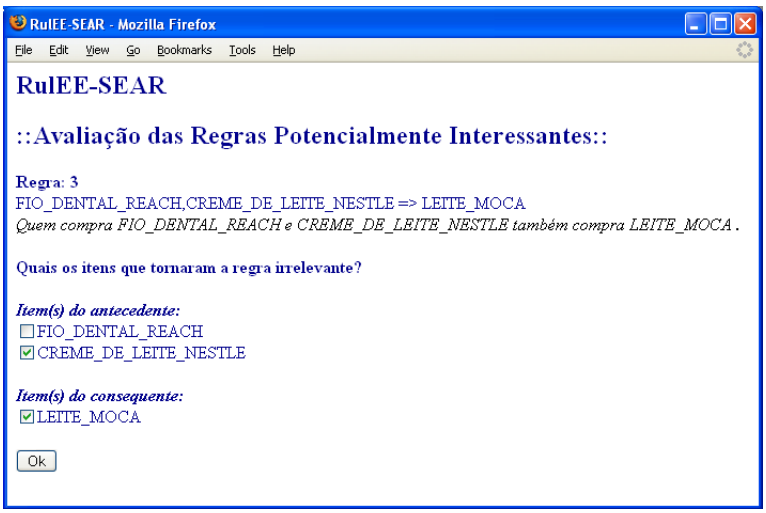


Figura 4.6: Eliminação das regras irrelevantes no RulEE-SEAR - Seleção dos itens irrelevantes

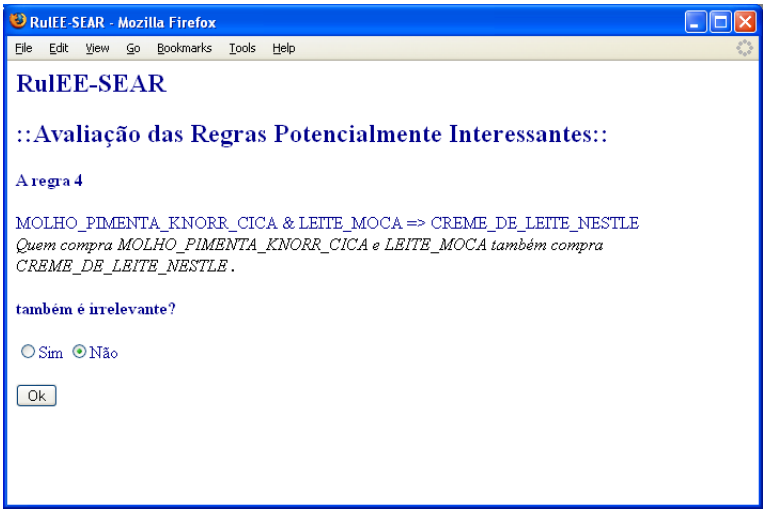


Figura 4.7: Eliminação das regras irrelevantes no RulEE-SEAR - Avaliação de regras possivelmente irrelevantes

do subconjunto RPI que foi selecionada, possuem os valores da medida indicada dentro do intervalo definido (Figura 4.10). Nessa fase o usuário pode editar as informações de cada regra, escrevendo comentários e marcando-as como interessantes ou não.

Durante a análise com medidas subjetivas o usuário pode visualizar as regras consideradas interessantes e finalizada a análise, o usuário pode encerrar o processo de exploração, que ficará disponível no RulEE-SEAR para futuras visualizações.

Em qualquer momento do processo o usuário pode sair do sistema e depois retornar à etapa que ainda não foi concluída. Os usuários também podem acessar os seus processos encerrados e os processos, encerrados e visíveis, de outros usuários do mesmo projeto. Para esses processos, o usuário pode visualizar as regras consideradas interessantes e as avaliações das regras potencialmente interessantes. Por meio dessas visualizações ocorre a disponibilização e compartilhamento do conhecimento entre os usuários do RulEE-SEAR.

RulEE-SEAR - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

RulEE-SEAR

::Análise das Regras de Associação::

Regras avaliadas como Conhecimento Inesperado / Útil:

	Regra	Inesperado	Útil	Interessante	Comentário
○	DETERGYPE & ACHOCNESCAU & CREME_DE_LEITE_NESTLE => LEITE_MOCA *Quem compra DETERGYPE e ACHOCNESCAU e CREME_DE_LEITE_NESTLE também compra LEITE_MOCA .*	Sim	Sim	Sim	
○	CHOCLACTA & ACHOCNESCAU & LEITE_MOCA & COCA_COLA => BISCNESTLE *Quem compra CHOCLACTA e ACHOCNESCAU e LEITE_MOCA e COCA_COLA também compra BISCNESTLE .*	Não	Sim	Sim	
○	DETERGLIMPOL & SPO_OMO & CREME_DE_LEITE_NESTLE & BISCNESTLE => ACUCAR_UNIAO *Quem compra DETERGLIMPOL e SPO_OMO e CREME_DE_LEITE_NESTLE e BISCNESTLE também compra ACUCAR_UNIAO .*	Sim	Sim	Sim	

Regras avaliadas como Conhecimento Óbvio do Dominio / Prévio:

	Regra	Óbvio	Prévio	Interessante	Comentário
○	FERMROYAL & FARTRIGO_RENATA & CREME_DE_LEITE_NESTLE => LEITE_MOCA *Quem compra FERMROYAL e FARTRIGO_RENATA e CREME_DE_LEITE_NESTLE também compra LEITE_MOCA .*	Não	Sim	Não	
○	ABSSLIVRE & MAIONESE_HELLMANN_S & CREME_DE_LEITE_NESTLE => LEITE_MOCA *Quem compra ABSSLIVRE e MAIONESE_HELLMANN_S e CREME_DE_LEITE_NESTLE também compra LEITE_MOCA .*	Não	Sim	Sim	
○	FARTRIGO_RENATA & BOMBRIL & CREME_DE_LEITE_NESTLE & COCA_COLA => ACUCAR_UNIAO *Quem compra FARTRIGO_RENATA e BOMBRIL e CREME_DE_LEITE_NESTLE e COCA_COLA também compra ACUCAR_UNIAO .*	Não	Sim	Sim	

Ver Regras Interessantes | Encerrar Exploração | Sair

Figura 4.8: Análise com medidas subjetivas no RulEE-SEAR - Apresentação das regras avaliadas

RulEE-SEAR - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

RulEE-SEAR

::Análise das Regras de Associação::

Regra 10 selecionada:

DETERGYPE & ACHOCNESCAU & CREME_DE_LEITE_NESTLE => LEITE_MOCA

Selecione uma medida:

- ◉ Conformidade (confmij)
- ○ Antecedente Inesperado (unexpCondij)
- ○ Conseqüente Inesperado (unexpConseqij)
- ○ Antecedente e Conseqüente Inesperados (bsUnexpij)

Defina um intervalo:

Valor mínimo:
0,3

Valor máximo:
1

Voltar - Regras Avaliadas | Selecionar Regras

Figura 4.9: Análise com medidas subjetivas no RulEE-SEAR - Seleção da medida subjetiva

RulEE-SEAR - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

RulEE-SEAR

::Análise das Regras de Associação::

Regra 10 selecionada:

Editar	Regra	Interessante	Comentário
○	DETERGYPE & ACHOCNESCAU & CREME_DE_LEITE_NESTLE => LEITE_MOCA *Quem compra DETERGYPE e ACHOCNESCAU e CREME_DE_LEITE_NESTLE também compra LEITE_MOCA .*	Sim	

Regras com confmij entre 0.3 e 1:

Editar	Regra	confmij	Interessante	Comentário
○	DETERGYPE & ARROZ_PRATO_FINO & CREME_DE_LEITE_NESTLE => LEITE_MOCA *Quem compra DETERGYPE e ARROZ_PRATO_FINO e CREME_DE_LEITE_NESTLE também compra LEITE_MOCA .*	0.6667	Sim	
○	MAIONESE_HELLMANN_S & ACHOCNESCAU & CREME_DE_LEITE_NESTLE & BISCNESTLE => LEITE_MOCA *Quem compra MAIONESE_HELLMANN_S e ACHOCNESCAU e CREME_DE_LEITE_NESTLE e BISCNESTLE também compra LEITE_MOCA .*	0.5000	Não	
○	BOMBRIL & ACHOCNESCAU & CREME_DE_LEITE_NESTLE & BISCNESTLE => LEITE_MOCA *Quem compra BOMBRIL e ACHOCNESCAU e CREME_DE_LEITE_NESTLE e BISCNESTLE também compra LEITE_MOCA .*	0.5000	Não	
○	FERMROYAL & CREME_DE_LEITE_NESTLE & ACUCAR_UNIAO => LEITE_MOCA *Quem compra FERMROYAL e CREME_DE_LEITE_NESTLE e ACUCAR_UNIAO também compra LEITE_MOCA .*	0.3333	Não	
○	FERMROYAL & CREME_DE_LEITE_NESTLE & COCA_COLA => LEITE_MOCA *Quem compra FERMROYAL e CREME_DE_LEITE_NESTLE e COCA_COLA também compra LEITE_MOCA .*	0.3333	Não	
○	MACRENATA & SAL_CISNE & CREME_DE_LEITE_NESTLE => LEITE_MOCA *Quem compra MACRENATA e SAL_CISNE e CREME_DE_LEITE_NESTLE também compra LEITE_MOCA .*	0.3333	Não	

Voltar

Figura 4.10: Análise com medidas subjetivas no RulEE-SEAR - Apresentação das regras do conjunto foco

4.5 Considerações Finais

Nesse capítulo foi apresentada a metodologia proposta para identificação de regras de associação interessantes combinando análises com medidas objetivas e subjetivas. Essa combinação é realizada com o objetivo de aproveitar as vantagens de cada tipo de medida e minimizar suas desvantagens. As medidas objetivas são mais gerais que as subjetivas e essa característica implica em vantagens e desvantagens para esse tipo de medida. Elas podem ser utilizadas independente do domínio da aplicação ou dos objetivos do processo de mineração de dados. No entanto, elas podem não ser suficientes para solucionar o problema da interessabilidade, visto que não consideraram fatores específicos da aplicação. Já as medidas subjetivas consideram esses fatores, utilizando o conhecimento e/ou interesses do usuário em seus cálculos, mas algumas limitações para o uso dessas medidas podem ser encontradas. O usuário pode ter dificuldade em fornecer seus interesses e/ou conhecimento, pois normalmente ele tem que quantificar conceitos imprecisos que possui e utilizar linguagens pré-definidas.

Nesse contexto, a metodologia proposta é composta por quatro etapas: análise objetiva, captação do conhecimento e dos interesses do especialista, processamento da avaliação e análise com medidas subjetivas. A análise objetiva segue o fluxo de análise proposto por Melanda (2004) e na análise com medidas subjetivas são utilizadas as medidas con-

formidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados, propostas por Liu, Hsu, Chen, & Ma (2000). Na metodologia proposta, o cálculo dessas medidas subjetivas é realizado de maneira diferenciada, não sendo necessário que o usuário forneça explicitamente seu conhecimento. O usuário apenas avalia algumas regras que são utilizadas como um conhecimento impreciso no cálculo das medidas.

Deve-se notar a integração que existe entre as etapas dessa metodologia. Uma etapa tem grande importância para a boa realização da próxima etapa e para atingir o objetivo da metodologia, ou seja, auxiliar o usuário na identificação de conhecimento interessante. A análise objetiva viabiliza a avaliação realizada na etapa de captação do conhecimento e dos interesses do especialista ao diminuir o número de regras a serem avaliadas, e a avaliação, por sua vez, realiza de maneira simples e fácil para o especialista a captação de seu conhecimento e interesses, necessária para o cálculo das medidas subjetivas.

Para viabilizar a aplicação da metodologia proposta foi desenvolvido um módulo de pós-processamento para o ambiente RulEE, o RulEE-SEAR, também descrito nesse capítulo. Assim, a primeira etapa da metodologia, a análise objetiva, pode ser realizada no $\mathcal{ARInE}$, um dos módulos do RulEE, e as demais etapas, no módulo de exploração de regras de associação com medidas subjetivas RulEE-SEAR. Utilizando esses dois módulos foi elaborado um estudo de caso com o objetivo de validar a metodologia proposta. No próximo capítulo é apresentado esse estudo de caso. Vale ressaltar que a metodologia foi validada no módulo RulEE-SEAR com as quatro medidas subjetivas citadas anteriormente, porém ela pode ser estendida com outras medidas subjetivas.

Capítulo

5

Estudo de Caso

A metodologia proposta no capítulo anterior visa auxiliar o usuário na identificação de regras de associação interessantes, combinando análises com medidas objetivas e subjetivas. Com o objetivo de validar a metodologia foi realizado um estudo de caso com uma base de dados sobre qualidade de vida urbana. O pós-processamento do conhecimento foi executado com a aplicação da metodologia proposta no ambiente RulEE, realizando a primeira etapa no $\mathcal{ARInE}$ e as demais no RulEE-SEAR. Durante todo o processo de mineração de dados houve a participação de um especialista do domínio, principalmente na etapa de pós-processamento, na qual foi realizada a captação do conhecimento e dos interesses do especialista (segunda etapa da metodologia). Neste capítulo é apresentado esse estudo de caso, sendo descritas todas as etapas do processo de mineração de dados realizadas, com ênfase no pós-processamento, etapa em que a metodologia está inserida.

5.1 Identificação do Problema

O processo de mineração de dados desse estudo de caso foi realizado com o objetivo de encontrar associações entre as características de determinadas regiões urbanas que interferem na qualidade de vida de seus residentes.

A base de dados utilizada, denominada QVU, é composta por um conjunto de variáveis de qualidade de vida consideradas na formulação de um indicador de qualidade de vida urbana (IQVU) de uma cidade de médio porte (São Carlos - SP) e pelo próprio índice. Essas variáveis são de natureza quantitativa e se referem às características e recursos pessoais dos habitantes da cidade, como grau de alfabetização e renda familiar, e às

características do lugar urbano, como infraestrutura e grau de acessibilidade espacial às atividades urbanas. Cada um dos 120 exemplos da base de dados QVU representa um setor urbano da cidade de São Carlos. Todos os 19 atributos são contínuos e não possuem valores ausentes. Na tabela 5.1 são apresentadas as descrições desses atributos.

Atributo	**Max**	**Min**	**Média**	**Desvio padrão**	**Descrição do atributo**
IQVU	-1,13	2,30	-0,13	0,53	Indicador de qualidade de vida urbana
acessibi	0,52	0,90	0,80	0,08	Nível de acessibilidade às atividades urbanas
agua_p	75,16	100	98,08	3,38	Percentual de domicílios atendidos pelo sistema de abastecimento de água
cmd_dom	3	12	5,70	1,21	Número médio de cômodos por domicílio
denscons	8566,67	782983,30	209655,55	121094,82	Densidade construída
densidad	2,15	157,62	61,52	35,23	Densidade populacional (hab./ha.)
dpi_p	0	4,17	0,17	0,52	Percentual de domicílios particulares improvisados
esgoto_p	71,20	100	95,58	4,71	Percentual de domicílios atendidos pelo sistema de coleta de esgoto sanitário
kilometers	0	4,79	2,04	0,96	Distância do n-ésimo setor ao setor central
lixo_p	75,16	100	98,00	3,46	Percentual de domicílios atendidos pelo sistema de coleta de resíduos sólidos
perocppr	44,62	97,52	68,56	9,88	Percentual de domicílios de propriedade do morador
pess_dom	2	5	3,29	0,54	Número médio de habitantes por domicílio
porpdimp	0	3,90	0,15	0,46	Percentual de habitantes em domicílios improvisados
porpdp	91,06	100	99,44	1,16	Percentual de habitantes em domicílios permanentes
porpopal	56,81	92,35	82,59	5,23	Percentual de indivíduos alfabetizados
rendchef	53467	885575	194930,50	117957,65	Renda média dos chefes de família (CR$)
rotas	9	87	40,27	23,14	Número de rotas de ônibus
sinstsan	0	19	0,72	2,12	Número de domicílios sem instalações sanitárias
stop_pto	20	471	105,73	81,21	Número de pontos de parada das rotas de ônibus

Tabela 5.1: Descrição dos atributos da base de dados QVU

5.2 Pré-Processamento

A primeira atividade realizada no pré-processamento dos dados foi a discretização dos 19 atributos. Durante essa etapa o especialista foi consultado sobre qual seria um bom número de intervalos para cada atributo. Com essa informação fornecida pelo especialista, a discretização dos atributos foi realizada na ferramenta Weka[1] (Witten & Frank, 2000), utilizando o método *equal-width binning*.

Após a discretização, os dados foram colocados no formato de entrada do algoritmo *Apriori*, utilizado na extração de padrões. Essa conversão de formato foi realizada com o *script hrd2set* fornecido juntamente com o *Apriori*. Assim, os dados no formato atributo-valor, como apresentado na Tabela 5.2, foram colocados no formato de itens-transações, como na Tabela 5.3.

	a_1	a_2	$\cdots$	a_m
e_1	x_{11}	x_{12}	$\cdots$	x_{1m}
e_2	x_{21}	x_{22}	$\cdots$	x_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
e_n	x_{n1}	x_{n2}	$\ldots$	x_{nm}

Tabela 5.2: Tabela atributo-valor

t_1	$a_1 = x_{11}$	$a_2 = x_{12}$	$\cdots$	$a_m = x_{1m}$
t_2	$a_1 = x_{21}$	$a_2 = x_{22}$	$\cdots$	$a_m = x_{2m}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_n	$a_1 = x_{n1}$	$a_2 = x_{n2}$	$\ldots$	$a_m = x_{nm}$

Tabela 5.3: Tabela de itens-transações

5.3 Extração de Padrões

Na etapa de extração de padrões foi utilizado o algoritmo *Apriori*[2], com os valores de suporte e confiança mínimos definidos como 20% e 50%, respectivamente. Foi definido um valor não muito alto para o suporte mínimo devido ao número relativamente pequeno de transações. Assim, com os valores utilizados, para que um *itemset* seja freqüente ele deve estar presente em, no mínimo, 24 transações (20% de 120) e, para que uma regra seja gerada com esse *itemset* ele deve estar presente em, pelo menos, metade das transações em que existem os itens que formam o antecedente.

Além dos valores de suporte e confiança mínimos, no algoritmo utilizado, também é possível configurar o número máximo de itens por regra. Nesse estudo de caso, esse número foi limitado a 3, visando facilitar a compreensão das regras.

[1] http://www.cs.waikato.ac.nz/ml/weka
[2] http://fuzzy.cs.uni-magdeburg.de/∼borgelt/apriori.html

Com essa configuração, o *Apriori* gerou 4.122 regras de associação, utilizando 30 itens freqüentes, ou seja, 30 itens com suporte maior ou igual ao suporte mínimo definido.

5.4 Pós-Processamento

O pós-processamento das regras obtidas foi realizado no ambiente RulEE. Para inserir as regras nesse ambiente foi necessário converter a saída do *Apriori* para a sintaxe padrão. Essa conversão foi realizada com um *script* do Discover (Melanda & Rezende, 2003).

A seguir são descritas as atividades realizas no pós-processamento seguindo a metodologia proposta no Capítulo 4. Essas atividades estão divididas em: análise objetiva, primeira etapa da metodologia, realizada no $\mathcal{ARInE}$, e exploração subjetiva, correspondendo às etapas de captação do conhecimento e dos interesses do especialista, processamento da avaliação e análise com medidas subjetivas, realizadas no RulEE-SEAR.

5.4.1 Análise Objetiva

Com as regras inseridas no RulEE, a primeira etapa da metodologia foi realizada no $\mathcal{ARInE}$, com as atividades definição do conjunto foco, análise de similaridade das medidas objetivas e definição do subconjunto RPI, descritas a seguir.

Definição do conjunto foco

Para definir o conjunto foco foram selecionadas todas as regras geradas na extração de padrões, visto que não havia interesse em um item específico. Assim, o conjunto foco foi formado com 4.122 regras.

Análise de similaridade das medidas objetivas

A fim de verificar a similaridade existente entre as 31 medidas objetivas cadastradas no $\mathcal{ARInE}$ foi utilizada a técnica de análise fatorial. Essa análise foi realizada na ferramenta estatística SPSS[3] com um arquivo, gerado no $\mathcal{ARInE}$, contendo os valores das medidas para 2.442 regras do conjunto foco (que tem no total 4.122 regras). Essas regras foram selecionadas no $\mathcal{ARInE}$ de maneira aleatória com o limite de 3.000 regras, máximo suportado pela ferramenta SPSS.

Com o resultado da análise fatorial, as 31 medidas foram distribuídas em quatro grupos, cada um correspondendo a um fator encontrado pela ferramenta. Essa distribuição é apresentada na Tabela 5.4. Duas das medidas, especificidade e convicção, foram coloca-

[3]http://www.spss.com/spss/

das em dois grupos, (2 e 3) e (3 e 1) respectivamente, pois os graus com que dois fatores descreviam cada medida eram muito próximos.

Medidas			
Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**
added value	cobertura	certainty factor	**confiança**
collective strength	cosine	convicção	erro
confiança negativa	especificidade	especificidade	laplace
confiança negativa relativa	IS	**especificidade relativa**	
confiança relativa	jaccard	satisfação	
convicção	sensibilidade	yules-Q	
gini index	**suporte**	yules-Y	
J_1			
kappa			
klosgen			
lift			
novidade			
odds ratio			
sensibilidade relativa			
ϕ-coefficient			
χ^2			

Tabela 5.4: Agrupamento das medidas objetivas obtido com a análise fatorial

Definição do subconjunto RPI

Com a análise de similaridade as medidas objetivas foram divididas em quatro grupos, de acordo com o comportamento de cada uma em uma amostra do conjunto de regras. Assim, para filtrar as regras do conjunto foco e definir o subconjunto RPI, foram selecionadas uma medida de cada grupo: lift (grupo 1), suporte (grupo 2), especificidade relativa (grupo 3) e confiança (grupo 4). Essas medidas foram selecionadas por pertencerem a apenas um dos grupos e devido à facilidade de interpretação de seus significados. Para verificar a distribuição dos valores dessas medidas no conjunto foco foram gerados gráficos de dispersão no $\mathcal{ARInE}$, apresentados nas Figuras 5.1, 5.2, 5.3 e 5.4. Em cada gráfico, para cada uma das regras do conjunto foco identificada pelo seu número no conjunto de regras (eixo x) existe um número (eixo y) referente ao valor da medida analisada.

As medidas suporte e confiança, como apresentado anteriormente, indicam a freqüência do *itemset* que forma a regra no conjunto de dados e a freqüência do *itemset* nas transações em que o antecedente ocorre. Os gráficos gerados para as medidas suporte e confiança são apresentados nas Figuras 5.1 e 5.2, respectivamente. Analisando a medida suporte, pode-se verificar que para grande parte das regras o valor dessa medida está próximo a $0,2$, valor mínimo definido para gerar as regras. Já no caso dos valores da medida confiança, a maioria das regras estão próximas ao valor máximo da medida. No entanto essas medidas não são adequadas para filtrar as regras, visto que foram utilizadas na extração de padrões. Todas as regras do conjunto foco possuem valores maiores ou iguais aos valores mínimos definidos, ou seja, valores aceitáveis para essas medidas de acordo com os objetivos do processo de mineração de dados.

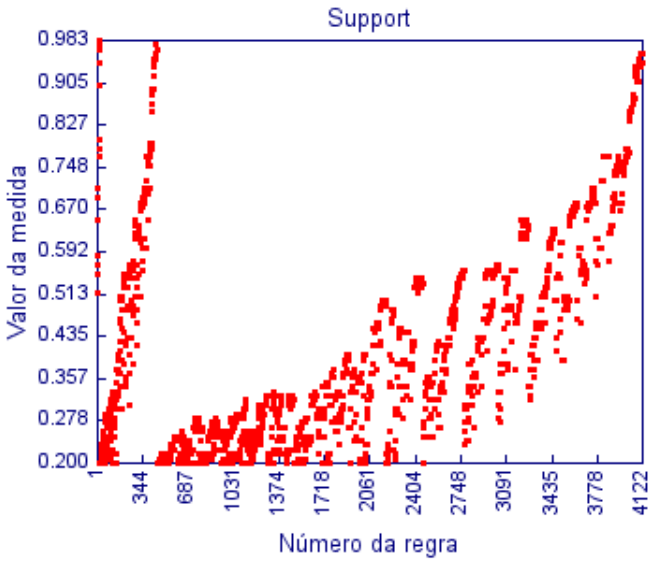


Figura 5.1: Gráfico de dispersão dos valores da medida suporte

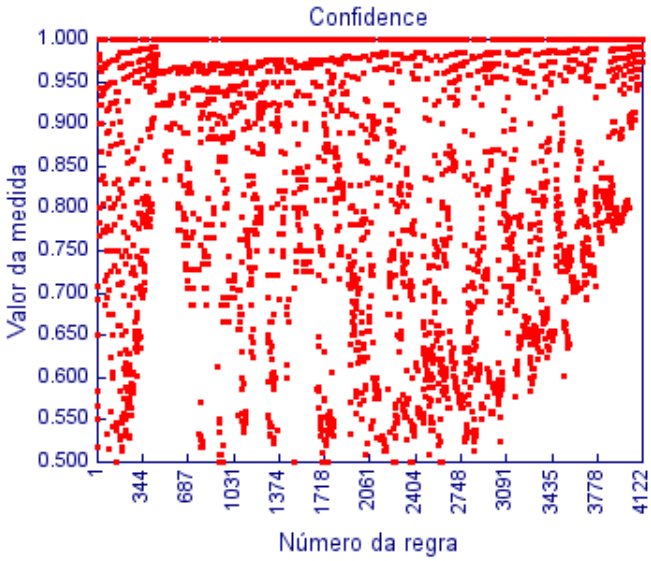


Figura 5.2: Gráfico de dispersão dos valores da medida confiança

A medida lift indica o quanto aumenta a freqüência do conseqüente na presença do antecedente. Assim, regras com lift igual a 1 possuem antecedente e conseqüente independentes, ou seja, a presença do antecedente não leva ao aumento ou à diminuição da ocorrência do conseqüente. No caso de regras com lift maior que 1, pode-se dizer que o antecedente influencia positivamente a freqüência do conseqüente e, no caso de lift menor que 1, o antecedente e o conseqüente são negativamente dependentes. Portanto, as regras interessantes devem possuir valores altos de lift. No gráfico da Figura 5.3 é apresentada a distribuição dessa medida para as regras do conjunto foco. Nesse gráfico pode ser verificado que poucas regras possuem valores altos para a medida lift. Assim, foi realizada no RuleE uma consulta ao conjunto foco selecionando as regras com lift maior que 2. As três regras encontradas, bem como seus valores de suporte (Sup), confiança (Conf), lift e especificidade relativa (RSpec), são apresentadas na Tabela 5.5.

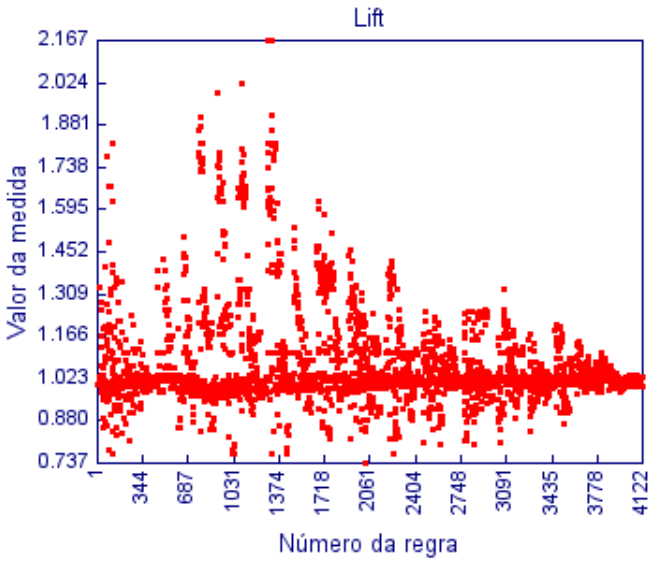


Figura 5.3: Gráfico de dispersão dos valores da medida lift

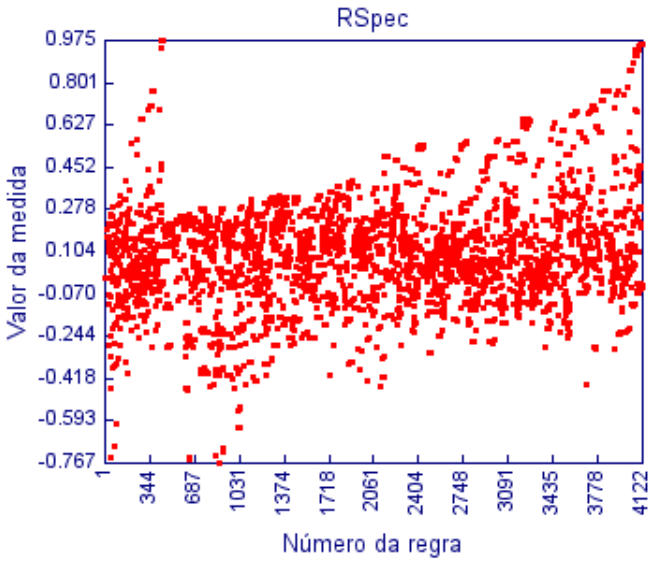


Figura 5.4: Gráfico de dispersão dos valores da medida especificidade relativa

Regra	Sup	Conf	Lift	RSpec
[R01] kilometers=(1.596667-3.193333], rotas=(-inf-35] ⇒ acessibi=(0.713-0.807]	0,2417	0,6744	2,0228	0,1833
[R02] cmd_dom=(5.25-7.5], acessibi=(0.807-inf) ⇒ kilometers=(-inf-1.596667]	0,2167	0,7223	2,1666	0,1750
[R03] acessibi=(0.807-inf), porpopal=(80.503333-inf) ⇒ kilometers=(-inf-1.596667]	0,3250	0,7222	2,1669	0,2625

Tabela 5.5: Regras obtidas com a restrição lift > 2

A medida especificidade relativa indica o ganho de especificidade da regra quando o conseqüente é acrescentado, ou seja, indica o grau de diminuição da ocorrência do antecedente sem o conseqüente. Assim, as regras interessantes devem possuir valores altos de especificidade relativa. A distribuição dos valores da medida especificidade relativa é apresentada no gráfico da Figura 5.4, no qual pode ser visto que poucas regras possuem os valores mais altos para essa medida (valores próximos a 1). Com uma consulta no conjunto

foco realizada no RulEE foram selecionadas 17 regras com especificidade relativa maior que $0,9$. Essas regras e os valores para as quatro medidas são apresentados na Tabela 5.6.

Regra	Sup	Conf	Lift	RSpec
[R04] agua_p=(91.719078-inf) ⇒ lixo_p=(87.578616-inf)	0,9417	1,0000	1,0169	0,9416
[R05] porpdimp=(-inf-1.30002] ⇒ dpi_p=(-inf-1.388889]	0,9750	1,0000	1,0256	0,9750
[R06] dpi_p=(-inf-1.388889] ⇒ porpdimp=(-inf-1.30002]	0,9750	1,0000	1,0256	0,9750
[R07] agua_p=(91.719078-inf), porpdimp=(-inf-1.30002] ⇒ porpdp=(97.018349-inf)	0,9167	1,0000	1,0344	0,9167
[R08] agua_p=(91.719078-inf), dpi_p=(-inf-1.388889] ⇒ porpdp=(97.018349-inf)	0,9167	1,0000	1,0344	0,9167
[R09] agua_p=(91.719078-inf), porpdp=(97.018349-inf) ⇒ lixo_p=(87.578616-inf)	0,9333	1,0000	1,0170	0,9333
[R10] agua_p=(91.719078-inf), porpdimp=(-inf-1.30002] ⇒ dpi_p=(-inf-1.388889]	0,9167	1,0000	1,0256	0.9167
[R11] agua_p=(91.719078-inf), dpi_p=(-inf-1.388889] ⇒ porpdimp=(-inf-1.30002]	0,9167	1,0000	1,0256	0,9167
[R12] agua_p=(91.719078-inf), porpdimp=(-inf-1.30002] ⇒ lixo_p=(87.578616-inf)	0,9167	1,0000	1,0169	0,9166
[R13] agua_p=(91.719078-inf), dpi_p=(-inf-1.388889] ⇒ lixo_p=(87.578616-inf)	0,9167	1,0000	1,0169	0,9166
[R14] agua_p=(91.719078-inf), sinstsan=(-inf-6.333333] ⇒ lixo_p=(87.578616-inf)	0,9333	1,0000	1,0170	0,9333
[R15] porpdp=(97.018349-inf), porpdimp=(-inf-1.30002] ⇒ dpi_p=(-inf-1.388889]	0,9500	1,0000	1,0256	0,9500
[R16] porpdp=(97.018349-inf), dpi_p=(-inf-1.388889] ⇒ porpdimp=(-inf-1.30002]	0,9500	1,0000	1,0256	0,9500
[R17] porpdimp=(-inf-1.30002], lixo_p=(87.578616-inf) ⇒ dpi_p=(-inf-1.388889]	0,9583	1,0000	1,0256	0,9583
[R18] dpi_p=(-inf-1.388889], lixo_p=(87.578616-inf) ⇒ porpdimp=(-inf-1.30002]	0,9583	1,0000	1,0256	0,9583
[R19] porpdimp=(-inf-1.30002], sinstsan=(-inf-6.333333] ⇒ dpi_p=(-inf-1.388889]	0,9583	1,0000	1,0256	0,9583
[R20] dpi_p=(-inf-1.388889], sinstsan=(-inf-6.333333] ⇒ porpdimp=(-inf-1.30002]	0,9583	1,0000	1,0256	0,9583

Tabela 5.6: Regras obtidas com a restrição especificidade relativa > 0,9

Analisando as Tabelas 5.5 e 5.6 pode-se verificar que as 20 regras selecionadas possuem valores aceitáveis para as quatro medidas analisadas. Todas as regras apresentam suporte $> 0,2$, confiança $> 0,5$, lift > 1 e especificidade relativa > 0. Dessa maneira, todas as regras foram consideradas potencialmente interessantes e foram utilizadas para formar o subconjunto RPI.

5.4.2 Exploração Subjetiva

Com o conjunto foco e o suconjunto RPI formados no $\mathcal{ARInE}$, foi realizado um processo de exploração subjetiva no RulEE-SEAR, compreendendo as etapas captação do conhecimento e dos interesses do especialista, processamento da avaliação e análise com medidas subjetivas, que são descritas a seguir.

Captação do conhecimento e dos interesses do especialista

Nessa etapa o especialista do domínio avaliou 7 regras do subconjunto RPI no RULEESEAR. Não foram avaliadas todas as 20 regras selecionadas porque 13 delas foram eliminadas por apresentarem itens considerados irrelevantes durante a avaliação. Na Tabela 5.7 são apresentadas as respostas do especialista para cada regra avaliada, apresentadas nas Tabelas 5.5 e 5.6. Nessa tabela, as colunas indicam o tipo de conhecimento que a regra pode representar (inesperado (Inesp.), útil, óbvio (Obv.), prévio (Prev.) e irrelevante (Irrel.)), além dos itens considerados irrelevantes e o interesse do especialista pela regra (Int.). As células marcadas indicam uma resposta afirmativa em relação à respectiva coluna.

Regra	**Inesp.**	**Útil**	**Obv.**	**Prev.**	**Irrel.**	**Itens Irrelevantes**	**Int.**
[R04]				■	■	agua_p=(91.719078-inf) e lixo_p=(87.578616-inf)	
[R05]			■		■	porpdimp=(-inf-1.30002] e dpi_p=(-inf-1.388889]	
[R01]	■						■
[R02]		■					■
[R03]		■					■
[R07]				■	■	porpdimp=(-inf-1.30002] e porpdp=(97.018349-inf)	
[R08]				■	■	dpi_p=(-inf-1.388889] e porpdp=(97.018349-inf)	

Tabela 5.7: Avaliação das regras do subconjunto RPI

Processamento da avaliação

O processamento da avaliação foi realizado com a eliminação das regras irrelevantes, paralelamente a avaliação, e com o cálculo das medidas subjetivas. A cada regra avaliada como conhecimento irrelevante foram eliminadas as regras do conjunto foco que apresentavam os itens indicados. A partir das 4 regras do subconjunto RPI avaliadas como irrelevantes (Tabela 5.7) foram eliminadas outras 272 regras do conjunto foco, sendo 13 regras pertencentes ao subconjunto RPI e que não chegaram a ser apresentadas para avaliação.

Assim, foram calculadas as quatro medidas subjetivas para cada regra do conjunto foco, não eliminada e não pertencente ao subconjunto RPI, em relação a cada uma das 3 regras avaliadas e não eliminadas.

Análise com medidas subjetivas

Com as medidas subjetivas calculadas, o especialista realizou uma análise das regras do conjunto foco. Nessa análise foram encontradas outras 5 regras interessantes, que não pertenciam ao subconjunto RPI. Essas regras são apresentadas na Tabela 5.8.

Regra	Sup	Conf	Lift	RSpec
cmd_dom=(5.25-7.5] ⇒ kilometers=(-inf-1.596667]	0,2167	0,5417	1,6249	0,1250
kilometers=(1.596667-3.193333], perocppr=(62.253333-79.886667] ⇒ acessibi=(0.713-0.807]	0,2500	0,6000	1,8000	0,1667
kilometers=(1.596667-3.193333], rendchef=(-inf-261494] ⇒ acessibi=(0.713-0.807]	0,2583	0,5345	1,6035	0,1458
kilometers=(1.596667-3.193333], rotas=(-inf-35] ⇒ cmd_dom=(-inf-5.25]	0,2667	0,7441	1,4402	0,1686
kilometers=(1.596667-3.193333], rotas=(-inf-35] ⇒ denscons=(-inf-266705.546667]	0,2667	0,7441	1,1448	0,0964

Tabela 5.8: Regras interessantes encontradas na análise com medidas subjetivas

Após essa análise o processo de exploração subjetiva no RulEE-SEAR foi encerrado e as 8 regras interessantes foram disponibilizadas ao especialista do domínio.

5.5 Análise dos Resultados

Com o estudo de caso realizado foi verificada a viabilidade da metodologia proposta. A aplicação das medidas objetivas reduziu o número de regras a serem avaliadas pelo especialista e a análise com as medidas subjetivas permitiu a identificação de regras interessantes não encontradas com as medidas objetivas.

Durante a avaliação o especialista notou que alguns atributos da base de dados eram correlacionados. Regras com esses atributos foram então consideradas irrelevantes. A identificação de 4 regras irrelevantes levou à eliminação de 276 regras do conjunto foco. A fim de verificar a validade da eliminação das regras, as regras do subconjunto RPI eliminadas automaticamente também foram apresentadas ao especialista. Analisando as 20 regras do subconjunto RPI o especialista identificou que elas representavam apenas 6 tipos de conhecimento, sendo que o RulEE-SEAR apresentou todos eles durante a avaliação, repetindo apenas um com a apresentação da sétima regra.

Vale ressaltar que as 3 regras consideradas interessantes durante a avaliação são as mesmas 3 regras selecionadas com a medida lift na análise objetiva e que todas as regras selecionadas com a medida especificidade relativa foram consideradas irrelevantes. Além disso, os valores da medida especificidade relativa para as regras interessantes identificadas na análise com medidas subjetivas (Tabela 5.8) são relativamente baixos enquanto os valores da medida lift são mais próximos ao limite estabelecido para essa medida durante a análise objetiva. Portanto, pode-se verificar que, para esse estudo de caso, a medida lift representou melhor o interesse do especialista, não destacando regras com atributos muito correlacionados.

5.6 Considerações Finais

Nesse capítulo foi descrito o estudo de caso realizado com o objetivo de mostrar a viabilidade da metodologia proposta no apoio à identificação de regras interessantes. O pós-processamento nesse estudo de caso foi iniciado com um conjunto com 4.122 regras geradas na extração de padrões. Na primeira etapa da metodologia foram selecionadas 20 regras com as medidas objetivas. Dessas regras, 3 foram consideradas interessantes pelo especialista do domínio e 276 regras foram eliminadas do conjunto foco por representarem conhecimento irrelevante. Posteriormente, na análise com medidas subjetivas, o especialista identificou outras 5 regras interessantes.

Com esse estudo de caso verificou-se que a metodologia proposta é viável e que a combinação de medidas objetivas e subjetivas pode auxiliar na identificação de regras de associação interessantes. No próximo capítulo são apresentadas as conclusões deste trabalho e propostas para trabalhos futuros.

Capítulo

6

Conclusão

Este trabalho teve início com um estudo do problema da interessabilidade de regras, em especial das regras de associação, abrangendo as técnicas e as ferramentas utilizadas para a identificação de conhecimento interessante com foco na aplicação de medidas de avaliação. O principal objetivo desse estudo foi investigar algumas medidas subjetivas e o uso da subjetividade do especialista do domínio na identificação de conhecimento interessante (primeiro objetivo específico deste trabalho, apresentado na Seção 1.2). Com esse estudo foram verificados os fatores que tornam uma regra de associação interessante e as limitações das medidas objetivas e subjetivas, temas abordados no Capítulo 3 e na Seção 4.1.

Após a identificação das vantagens e desvantagens da aplicação de medidas objetivas e subjetivas foi definida a hipótese que passou a guiar o desenvolvimento deste trabalho. Essa hipótese afirma que, visando a identificação de conhecimento interessante a determinado especialista, é possível utilizar as medidas objetivas e subjetivas de maneira combinada e aproveitar as vantagens de cada tipo de medida. A principal vantagem das medidas objetivas é que elas são mais gerais, pois no cálculo dessas medidas são considerados apenas a estrutura das regras e os dados utilizados na extração de padrões. Por outro lado, essa característica das medidas objetivas implica, da mesma maneira, em uma desvantagem, visto que a interessabilidade de uma regra também depende muito do objetivo e do conhecimento do especialista. Uma regra interessante a um especialista em determinado momento pode não ser interessante em outro momento ou a outro especialista. Por considerarem fatores específicos do domínio ou dos usuários, as medidas subjetivas podem expressar essas diferenças. No entanto, a maneira como o especialista deve fornecer seus interesses e conhecimento pode resultar em uma desvantagem no uso dessas medidas, pois em geral o especialista tem que utilizar uma linguagem pré-definida

e/ou quantificar conceitos imprecisos que possui.

Considerando a hipótese definida, foi especificada a metodologia para identificação de regras de associação interessantes apresentada no Capítulo 4. A primeira etapa dessa metodologia compreende uma análise objetiva, com consultas e aplicação de medidas objetivas. Por serem mais gerais, as medidas objetivas são utilizadas como um filtro, para selecionar um subconjunto de regras potencialmente interessantes. Essas regras são apresentadas ao especialista, que as avalia de acordo com seus interesses e conhecimento. Com essa avaliação é feita a captação da subjetividade do especialista, de maneira fácil para o mesmo, visto que ele não precisa explicitar seu conhecimento, devendo apenas avaliar as regras marcando uma ou mais opções que indicam o tipo de conhecimento que cada uma representa. Considerando as regras avaliadas como um conhecimento fornecido pelo usuário é possível calcular as medidas subjetivas conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados. Essas medidas podem então ser utilizadas na última etapa da metodologia, guiando o usuário na análise das regras.

A metodologia proposta abrange os dois fatores que podem tornar uma regra de associação interessante da seguinte maneira: por meio das consultas, na análise objetiva, é possível selecionar as regras que apresentam os itens que são interessantes ao especialista e com as medidas subjetivas é possível identificar associações interessantes entre itens. Com o desenvolvimento dessa metodologia foi atingido o segundo objetivo específico e o primeiro objetivo geral deste trabalho (apresentados na Seção 1.2).

Para viabilizar a aplicação da metodologia proposta foi desenvolvido um módulo de exploração de regras de associação com medidas subjetivas para o ambiente RulEE, o RulEE-SEAR, apresentado na Seção 4.4. Assim, todas as etapas da metodologia podem ser realizadas no ambiente RulEE. A primeira etapa, a análise objetiva, pode ser realizada no $\mathcal{ARInE}$ e as demais etapas, no RulEE-SEAR. Além de possibilitar o uso de medidas subjetivas no RulEE, o RulEE-SEAR também permite a tradução de regras de associação, visando facilitar a participação do usuário na análise das regras. Com o desenvolvimento desse módulo foram cumpridos o terceiro objetivo geral e, o terceiro e o quarto objetivos específicos deste trabalho, dado que além de permitir o uso das medidas subjetivas, o RulEE-SEAR também possibilita o compartilhamento de conhecimento entre usuários de um mesmo projeto por meio das visualizações disponíveis.

Após o desenvolvimento do módulo, com o objetivo de validar a metodologia proposta foi realizado um estudo de caso, descrito no Capítulo 5, com a participação de um especialista em qualidade de vida urbana. Com esse estudo de caso verificou-se que com a aplicação da metodologia proposta é possível identificar regras de associação interessantes aproveitando as vantagens de cada tipo de medida e facilitando a participação do especialista do domínio. Assim, foram cumpridos o segundo objetivo geral e o quinto objetivo específico deste trabalho.

Portanto, foram atingidos todos os objetivos, gerais e específicos, deste trabalho e a hipótese foi validada. Como principais contribuições pode-se destacar:

- Proposta de uma metodologia para identificação de regras de associação utilizando medidas objetivas e subjetivas de maneira a aproveitar as vantagens de cada uma e facilitar a participação do especialista do domínio nesse processo.
- Desenvolvimento de um módulo para o ambiente RulEE que possibilita o cálculo e a aplicação de medidas subjetivas para a identificação de regras de associação interessantes.

Visando complementar o trabalho relatado nesta dissertação e trazer novas contribuições para o uso da subjetividade do usuário na identificação de regras de associação interessantes, a seguir são apresentadas algumas propostas para trabalhos futuros.

- Incorporação de outras medidas subjetivas à metodologia proposta e ao RulEE-SEAR, que no momento suportam apenas as medidas conformidade, antecedente inesperado, conseqüente inesperado e antecedente inesperados.
- Expansão do RulEE-SEAR para que possam ser utilizadas taxonomias e regras generalizadas.
- Realização de mais experimentos utilizando a metodologia proposta, com outros especialistas do domínio utilizado no estudo de caso já realizado ou com bases de dados reais e especialistas de outros domínios.
- Criação de uma base de conhecimento com as avaliações realizadas pelos especialistas de determinado domínio e sua utilização para eliminação automática de regras que representam conhecimento óbvio do domínio no início do pós-processamento.
- Utilização de técnicas de visualização de informações no RulEE-SEAR, para auxiliar a análise das regras com as medidas subjetivas, por exemplo, apresentando ao especialista as regras em um mapa, sendo que a distância de uma regra do conjunto foco a uma regra avaliada represente o valor de determinada medida.

Referências Bibliográficas

Adamo, J.-M. (2001). *Data Mining for Association Rules and Sequential Patterns*. Springer-Verlag. 29

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. 7, 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. 15, 16, 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. 2, 15, 17

Antonie, M. & O. Zaane (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 [08/11/2005]. 7, 15

Baesens, B., S. Viaene, & J. Vanthienen (2000). Post-processing of association rules. In *Proceedings of the Workshop on Post-Processing in Machine Learning and Data Mining, The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2000)*, Boston, MA, USA, pp. 2–8. 15, 27

Batista, G. E. A. P. A., A. C. P. L. F. Carvalho, & M. C. Monard (2000). Applying one-sided selection to unbalanced datasets. In *Proceedings of the Mexican International Conference on Artificial Intelligence – MICAI'00*, Lecture Notes in Artificial Intelligence, pp. 315–325. Springer-Verlag. 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. 7, 15

Brin, S., R. Motwani, J. D. Ullman, & S. Tsur (1997). Dynamic itemset counting and implication rules for market basket data. In J. Peckham (Ed.), *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, pp. 255–264. 17, 28, 29, 30

Bruha, I. & A. Famili (2000). Postprocessing in machine learning and data mining. *ACM SIGKDD Explorations Newsletter 2*(2), 110–114. 13

Carvalho, D. R., A. A. Freitas, & N. F. F. Ebecken (2003). A critical review of rule surprisingness measures. In *Proceedings of the Fourth International Conference on Data Mining - Data Mining IV*, pp. 545–556. WIT Press. 27

Clementini, E., P. D. Felice, & K. Koperski (2000). Mining multiple-level spatial association rules for objects with a broad boundary. *Data & Knowledge Engineering 34*, 251–270. 15

Domingues, M. A. (2004). Generalização de regras de associação. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. 27, 36

Fayyad, U. M., G. Piatetsky-Shapiro, & P. Smyth (1996). From data mining to knowledge discovery: An overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, Capítulo 1, pp. 1–34. AAAI Press. 8

Freitas, A. A. (1998). On objective measures of rule surprisingness. In J. M. Zytkow & M. Quafafou (Eds.), *Proceedings of the 2nd European Conference on the Principles of Data Mining and Knowledge Discovery*, Volume 1510 of *Lecture Notes in Artificial Intelligence*, pp. 1–9. 23

Freitas, A. A. (1999). On rule interestingness measures. *Knowledge-Based Systems 12*(5–6), 309–315. 24

Glymour, C., D. Madigan, D. Pregibon, & P. Smyth (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery 1*(1), 11–28. 11

Han, J., J. Pei, & Y. Yin (2000). Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of data*, pp. 1–12. ACM Press. 17

Hilderman, R. J. & H. J. Hamilton (2001). Evaluation of interestingness measures for ranking discovered knowledge. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD'01*, pp. 247–259. 9, 22

Hipp, J., U. Güntzer, & G. Nakhaeizadeh (2002). Data mining of association rules and the process of knowledge discovery in databases. In *Advances in Data Mining*, Volume 2394 of *Lecture Notes in Artificial Intelligence*, pp. 15–36. 7, 15

Hipp, J., C. Mangold, U. Guntzer, & G. Nakhaeizadeh (2002). Efficient rule retrieval and postponed restrict operations for association rule mining. In M.-S. Cheng, P. S. Yu, & B. Liu (Eds.), *6th Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD'02*, Volume 2336 of *LNCS*, pp. 52–65. 27

Houtsma, M. & A. Swami (1995). Set-oriented mining for association rules in relational databases. In P. S. Yu & A. L. P. Chen (Eds.), *Proceedings of the 11th International Conference on Data Engineering*, pp. 25–33. 17

Hussain, F., H. Liu, & H. Lu (2000). Relative measure for mining interesting rules. In *4th European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD'00 – Workshop in Knowledge Management: Theory and Applications*. Disponível em: http://chirouble.univ-lyon2.fr/∼pkdd2000/Download/WS5_09.pdf [24/09/2004]. 25

Imieliński, T. & A. Virmani (1998). Association rules... and what's next? – towards second generation data mining systems. In W. Litwin, T. Morzy, & G. Vossen (Eds.), *Advances in Databases and Information Systems, Proceedings of the 2nd East European Symposium ADBIS'98*, Volume 1475 of *LNCS*, pp. 6–25. 18

Jorge, A. (2004). Hierarchical clustering for thematic browsing and summarization of large sets of association rules. In *Proceedings of the 4th SIAM International Conference on Data Mining (SIAM'04)*. Disponível em: http://www.niaad.liacc.up.pt/∼amjorge/docs/SDM04-Jorge.pdf [08/11/2005]. 27

Jorge, A., J. Poças, & P. Azevedo (2002). A post-processing environment for browsing large sets of association rules. In *Proceedings of Second International Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, pp. 53–64. 2, 27, 34

Klemettinen, M., H. Mannila, P. Ronkainen, H. Toivonen, & A. I. Verkamo (1994). Finding interesting rules from large sets of discovered association rules. In N. R. Adam, B. K. Bhargava, & Y. Yesha (Eds.), *Proceedings of the Third International Conference on Information and Knowledge Management CIKM'94*, pp. 401–407. ACM Press. 22, 25

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

Lavrač, N., P. Flach, & R. Zupan (1999). Rule evaluation measures: A unifying view. In S. Dzeroski & P. Flach (Eds.), *Proceedings of the 9th International Workshop on Inductive Logic Programming ILP'99*, Volume 1634 of *Lecture Notes in Artificial Intelligence*, pp. 174–185. Springer-Verlag. 2, 23, 27, 28, 29

Liu, B. & W. Hsu (1996). Post-analysis of learned rules. In *Proceedings of the 13th National Conference on Artificial Intelligence AAAI'96*, pp. 828–834. 9, 25, 26

Liu, B., W. Hsu, & S. Chen (1997). Using general impressions to analyze discovered classification rules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining KDD-97*, pp. 31–36. 25

Liu, B., W. Hsu, S. Chen, & Y. Ma (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems & their Applications 15*(5), 47–55. 22, 25, 27, 28, 31, 33, 34, 40, 41, 48, 49, 64, 91

Liu, B., W. Hsu, L.-F. Mun, & H.-Y. Lee (1996). Finding interesting patterns using user expectations. Relatório Técnico TRA7/96, Department of Information Systems and Computer Science, National University of Singapore. Disponível em: http://citeseer.ist.psu.edu/liu96finding.html [28/10/2005]. 23, 26

Ma, Y., B. Liu, & C. K. Wong (2000). Web for data mining: Organizing and interpreting the discovered rules using the web. *SIGKDD Explorations 2*(1), 16–23. 34

Ma, Y., C. K. Wong, & B. Liu (2000). Effective browsing of the discovered association rules using the web. In *Proceedings of the Workshop on Post-Processing in Machine Learning and Data Mining, The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA. Disponível em: http://www.cs.uic.edu/∼liub/publications/papers_topics.html [01/11/2005]. 34

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. 8

Melanda, E. A. (2004). *Pós-processamento de Regras de Associação.* Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. 2, 3, 4, 18, 26, 27, 30, 34, 37, 44, 45, 49, 63, 90

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/rel_tec/RT_206.pdf [28/10/2005]. 36, 68

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. 7, 15

Padmanabhan, B. & A. Tuzhilin (2000). Small is beautiful: Discovering the minimal set of unexpected patterns. In *Proceedings of the International Conference in Knowledge Discover and Data Mining (KDD-00)*, pp. 54–63. 2, 22

Park, J. S., M.-S. Chen, & P. S. Yu (1997). Using a hash-based method with transaction trimming for mining association rules. *IEEE Transactions on Knowledge and Data Engineering 9*(5), 813–825. 17

Paula, M. F. (2003). Ambiente para exploração de regras. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. 3, 18, 27, 34, 91

Pazzani, M. J. (2000). Knowledge discovery from data? *IEEE Intelligent Systems 15*(2), 10–13. 1

Pei, J., J. Han, & R. Mao (2000). CLOSET: An efficient algorithm for mining frequent closed itemsets. In D. Gunopulos & R. Rastogi (Eds.), *Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 21–30. 17

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pp. 229–248. AAAI/MIT Press. 23

Piatetsky-Shapiro, G. & C. J. Matheus (1994). The interestingness of deviations. In *Proceedings of the Knowledge Discovery in Databases KDD'94*, pp. 23–36. 10, 22

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/rel_tec/RT_145.pdf [28/10/2005]. 36

Pugliesi, J. B. (2004). *Pós-processamento de Regras de Regressão.* Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. 14

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/rel_tec/RT_193.zip [28/01/2004]. 36

Rathjens, D. (1996). *MineSet User's Guide.* Silicon Graphics, Inc. 28, 29

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. 18

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. 1, 8, 9

Sahar, S. (1999). Interestingness via what is not interesting. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pp. 332–336. ACM Press. 25, 26

Semenova, T., M. Hegland, W. Graco, & G. Williams (2001). Effectiveness of mining association rules for identifying trends in large health databases. In F. J. Kurfess & M. Hilario (Eds.), *Workshop on Integrating Data Mining and Knowledge Management, The 2001 IEEE International Conference on Data Mining – ICDM'01*, San Jose, California, USA. Disponível em: http://cui.unige.ch/∼hilario/icdm-01/DM-KM-Final/Semenova.pdf [22/10/2004]. 15

Shekar, B. & R. Natarajan (2002). A fuzzy-graph-based approach to the determination of interestingness of association rules. In D. Karagiannis & U. Reimer (Eds.), *Proceedings of the 4th International Conference on Practical Aspects of Knowledge Management – PAKM 2002*, Volume 2569 of *Lecture Notes in Computer Science*, pp. 377–388. Springer. 25

Shekar, B. & R. Natarajan (2004). A framework for evaluating knowledge-based interestingness of association rules. *Fuzzy Optimization and Decision Making 3*(2), 157–185. 25

Silberschatz, A. & A. Tuzhilin (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering 8*(6), 970–974. 2, 23, 25, 30, 31

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. 53

Srikant, R. & R. Agrawal (1997). Mining generalized association rules. *Future Generation Computer Systems 13*(2/3), 161–180. 27

Tan, P. N., V. Kumar, & J. Srivastava (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 32–41. 22

Wang, K., Y. Jiang, & L. V. S. Lakshmanan (2003). Mining unexpected rules by pushing user dynamics. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD'03*, pp. 246–255. ACM Press. 25

Wang, K., L. Tang, J. Han, & J. Liu (2002). Top down fp-growth for association rule mining. In *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 334–340. Springer-Verlag. 17

Wang, K., S. H. W. Tay, & B. Liu (1998). Interestingness-based interval merger for numeric association rules. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, New York City, New York, USA, pp. 121–128. 29

Webb, G. I. (1995). OPUS: An efficient admissible algorithm for unordered search. *Journal of Artificial Intelligence Research 3*, 431–465. 17

Weiss, S. M. & N. Indurkhya (1998). *Predictive Data Mining: A Pratical Guide.* Morgan Kaufmann Publishers Inc. 11, 12

Witten, I. H. & E. Frank (2000). *Data Mining: practical machine learning tools and techiniques with Java implementations.* Morgan Kaufmann. 67

Zaki, M. & C. Hsiao (2002). Charm: An efficient algorithm for closed itemset mining. In R. Grossman, J. Han, V. Kumar, H. Mannila, & R. Motwani (Eds.), *2nd SIAM International Conference on Data Mining.* Disponível em: http://www.siam.org/meetings/sdm02/proceedings/sdm02-27.pdf [28/10/2005]. 17

Zhang, C. & S. Zhang (2002). *Association Rule Mining: Models and Algorithms*, Volume 2307 of *Lecture Notes in Artificial Intelligence.* Springer-Verlag New York, Inc. Disponível em: http://www.springer.com/sgw/cda/frontpage/0,,5-164-22-2209306-0,00.html [28/10/2005]. 16

Zheng, C. & Y. Zhao (2003). A distance-based approach to find interesting patterns. In *Proceedings of 5th International Conference on Data Warehousing and Knowledge Discovery – DaWaK 2003*, Volume 2737 of *Lecture Notes in Computer Science*, pp. 299–308. Springer-Verlag. 25

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. 12

Zuffo, J. A. (2002). *A Sociedade e a Economia no Novo Milênio - Os empregos e as empresas no turbulento alvorecer do século XXI.* Editora Manole. 1

Apêndice

A

Documento de Requisitos do RulEE-SEAR

A.1 Introdução

O propósito deste documento é definir os requisitos do módulo de exploração de regras de associação com medidas subjetivas do ambiente RulEE, denominado RulEE-SEAR. O objetivo principal desse módulo é apoiar os usuários na identificação de regras de associação interessantes por meio de uma avaliação subjetiva das regras e posterior análise com medidas subjetivas.

Assim, este documento está dividido em três seções. Na Seção A.1 é apresentada uma breve introdução sobre o conteúdo deste documento. Uma descrição geral do RulEE-SEAR é apresentada na Seção A.2 e, na Seção A.3, são descritos os seus requisitos específicos.

A.1.1 Escopo

O RulEE-SEAR, módulo de exploração de regras de associação com medidas subjetivas do ambiente RulEE, possibilita a identificação de regras de associação interessantes utilizando o conhecimento de um especialista do domínio, que deve participar desse processo de avaliação.

Antes de iniciar o processo de exploração subjetiva nesse módulo, as regras de associação extraídas durante a etapa de extração de padrões devem ser inseridas no ambiente RulEE e analisadas objetivamente no ambiente $\mathcal{ARInE}$ (*Environment for Association Rules Interactive Exploration*).

O RulEE-SEAR utiliza dois subconjuntos do conjunto de regras resultante da etapa

de extração de padrões, chamados de conjunto foco e subconjunto de regras potencialmente interessantes. Esses subconjuntos são formados durante a análise objetiva realizada no *ARInE*. O conjunto foco é um subconjunto do conjunto de regras de associação ao qual pertencem as regras que apresentam os itens interessantes ao usuário. Já o subconjunto de regras potencialmente interessantes é um subconjunto do conjunto foco selecionado por meio do uso de uma ou mais medidas objetivas como filtro. Dados esses dois subconjuntos, o módulo permite que o usuário especialista do domínio avalie as regras do subconjunto de regras potencialmente interessantes por meio de um questionário e utiliza essa avaliação para calcular o valor das medidas subjetivas conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados, cadastradas no RulEE. Após o cálculo dessas medidas o usuário pode utilizá-las para analisar as regras do conjunto foco, e assim, identificar as regras interessantes.

Durante as etapas de avaliação e análise das regras, o RulEE-SEAR permite que o usuário salve as regras que achar interessante e escreva um comentário sobre elas. Essas regras salvas e os respectivos comentários, bem como as respostas dadas pelo especialista do domínio durante a avaliação das regras do subconjunto de regras potencialmente interessantes, podem ficar disponíveis a outros usuários do mesmo projeto, cadastrado no RulEE.

A fim de facilitar a compreensão das regras, o RulEE-SEAR possibilita a definição de traduções para as regras de associação e para seus itens, que são apresentadas juntamente com as regras durante a avaliação e a análise com medidas subjetivas.

A.1.2 Definições, Acrônimos e Abreviações

Nessa seção são dadas as definições, acrônimos e abreviações de termos que devem ser considerados no módulo de exploração de regras de associação com medidas subjetivas.

ARInE *Environment for Association Rules Interactive Exploration.* Ambiente que permite a análise de regras de associação baseada em consultas e medidas objetivas (Melanda, 2004).

Análise subjetiva de regras Análise das regras utilizando medidas subjetivas como guia.

Avaliação subjetiva de regras avaliação de um conjunto de regras realizada por um especialista do domínio, por meio de um questionário.

Conjunto foco Subconjunto do conjunto de regras de associação obtido na etapa de extração de padrões formado pelas regras que apresentam os itens que são interessantes para o especialista do domínio ou usuário final.

Medidas Subjetivas As medidas subjetivas utilizadas no RulEE-SEAR são conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados, definidas por Liu et al. (2000).

Processo de exploração subjetiva Compreende as etapas de avaliação subjetiva e análise com medidas subjetivas, calculadas com base na avaliação.

Projeto Conjuntos de regras de um mesmo domínio ou com objetivos em comum. Cada projeto pode ter mais de um usuário cadastrado.

RulEE *Rule Exploration Environment.* Ambiente para exploração e disponibilização de regras de classificação, regressão e associação (Paula, 2003).

Subconjunto de regras interessantes Subconjunto de um conjunto foco, formado por regras que são de interesse para o usuário que realizou a exploração.

Subconjunto de regras potencialmente interessantes Subconjunto de um conjunto foco, selecionado utilizando-se uma ou mais medidas objetivas como filtro durante uma análise realizada no $\mathcal{ARInE}$.

A.2 Descrição Geral do RulEE-SEAR

O objetivo do módulo RulEE-SEAR é auxiliar a identificação de regras de associação interessantes, utilizando o conhecimento que o especialista tem sobre o domínio. Portanto, a participação do especialista nesse processo é de extrema importância. Para facilitar essa participação, o RulEE-SEAR possibilita que as regras de associação sejam traduzidas, bem como os itens que as formam.

O processo de exploração subjetiva nesse ambiente é composto por duas etapas: avaliação e análise subjetivas. Durante a etapa de avaliação subjetiva, o especialista fornece o seu interesse e o conhecimento que tem sobre o domínio ao responder um questionário avaliando as regras do subconjunto de regras potencialmente interessantes. A partir dessa avaliação, o módulo calcula o valor das medidas subjetivas conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados, cadastradas no RulEE, para as regras do conjunto foco. Com esses valores disponíveis, inicia-se a etapa de análise subjetiva. Nessa etapa, o usuário pode navegar pelas regras do conjunto foco a partir das regras avaliadas, utilizando as medidas subjetivas como guia.

Tanto na etapa de avaliação quanto na de análise, ao encontrar uma regra interessante o especialista pode salvar essa regra e escrever um comentário sobre ela, formando assim, um subconjunto de regras interessantes.

No RulEE-SEAR, um usuário também pode acessar as explorações realizadas por outros usuários do projeto que ele é integrante, desde que o usuário responsável pela exploração permita. Assim, os diferentes usuários podem compartilhar suas avaliações.

A.2.1 Funções do Módulo

Para atingir seu objetivo de auxiliar a identificação de regras de associação interessantes inseridas no RulEE, o RulEE-SEAR deve possuir as seguintes funções:

- Criação, manipulação e finalização de um processo de exploração subjetiva;
- Avaliação de um subconjunto de regras potencialmente interessantes;
- Eliminação de regras irrelevantes;
- Cálculo de medidas subjetivas;
- Análise com medidas subjetivas das regras de um conjunto foco;
- Visualização de explorações realizadas;
- Definição de traduções para um conjunto de regras.

A.2.2 Características dos Usuários

Os usuários do módulo de exploração de regras de associação com medidas subjetivas são os usuários do ambiente RulEE, que são classificados em:

Administrador usuário responsável pelo RulEE, capaz de realizar todas as funções do ambiente.

Analista usuário especialista no processo de mineração de dados e responsável por sua execução.

Especialista do domínio usuário que possui amplo conhecimento do domínio da aplicação e fornece apoio para a execução do processo.

Usuário final usuário que utiliza o conhecimento extraído na mineração de dados para auxiliá-lo em um processo de tomada de decisão.

Como o módulo não possui nenhuma funcionalidade administrativa, todas as funcionalidades podem ser realizadas por qualquer um dos usuários do RulEE.

A.2.3 Suposições e Dependências

Para realizar uma exploração subjetiva de um conjunto de regras, este deve ser previamente inserido no RulEE e ser realizada sua análise objetiva no $\mathcal{ARInE}$. Com essa análise objetiva são formados o conjunto foco e o subconjunto de regras potencialmente interessantes, que são necessários para a exploração subjetiva realizada no módulo.

As medidas subjetivas conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados devem estar cadastradas no RulEE.

A.3 Requisitos Específicos

A.3.1 Requisitos Funcionais

Os requisitos funcionais do RulEE-SEAR estão organizados de acordo com as funções do módulo, apresentadas na Seção A.2.

1. **Requisitos para criação, manipulação e finalização de um processo de exploração**

   (a) O módulo deve permitir que qualquer usuário crie um novo processo de exploração subjetiva com um conjunto de regras de um projeto que ele é membro.

   (b) No momento da criação de um novo processo de exploração subjetiva, o módulo deve apresentar ao usuário os conjuntos de regras do projeto selecionado, e respectivos conjuntos focos e subconjuntos de regras potencialmente interessantes, que ele pode utilizar no processo que está sendo criado.

   (c) O módulo deve solicitar os seguintes dados para a criação de um novo processo de exploração subjetiva: conjunto foco, conjunto de regras potencialmente interessantes, descrição do processo, objetivo e interesse do usuário que está realizando o processo e se o processo pode ficar visível a outros usuários do mesmo projeto.

   (d) O módulo deve armazenar para cada processo de exploração subjetiva os dados solicitados, além do código do usuário e código do processo, que deve servir para identificar unicamente cada processo.

   (e) O módulo deve permitir que um usuário dê continuidade a um processo que ele tenha iniciado anteriormente, sendo identificado por seu código e descrição, desde que o processo ainda não tenha sido encerrado.

   (f) Caso o usuário queira dar continuidade a um processo de exploração iniciado anteriormente, o módulo deve verificar quais etapas já foram realizadas e iniciar na etapa que estiver pendente.

   (g) O módulo deve permitir que o usuário encerre um processo de exploração que tenha sido criado por ele em qualquer etapa do processo.

2. **Requisitos para avaliação subjetiva de um subconjunto de regras e eliminação de regras irrelevantes**

(a) O módulo deve permitir que um usuário avalie as regras do subconjunto de regras potencialmente interessantes indicado durante a criação do processo de exploração subjetiva em questão.

(b) O módulo deve apresentar ao usuário as regras do subconjunto de regras potencialmente interessantes individualmente e ordenadas pelo tamanho do *itemset*, sendo que as menores devem ser apresentadas antes das maiores.

(c) Para cada regra, o módulo deve apresentar a sua tradução, se existir, e as opções de avaliação: representa um conhecimento inesperado, representa um conhecimento útil, representa um conhecimento óbvio, representa um conhecimento prévio, representa um conhecimento irrelevante.

(d) O módulo deve permitir que o usuário selecione uma ou mais opções de avaliação para cada regra.

(e) Para as regras que forem selecionadas a opção representa um conhecimento irrelevante, o módulo deve solicitar do usuário os itens que tornam a regra irrelevante (caso a regra seja formada por mais de dois itens).

(f) O módulo deve eliminar as regras do subconjunto de regras potencialmente interessantes que foram avaliadas como conhecimento irrelevante.

(g) O módulo deve eliminar as regras do conjunto foco que apresentam os itens que tornam uma regra irrelevante indicados pelo usuário durante a avaliação na mesma posição em que são encontrados na regra avaliada.

(h) O módulo deve verificar se existem regras no conjunto foco com os itens que tornam uma regra irrelevante em posições distintas e, em caso positivo, questionar o usuário se essas regras também devem ser eliminadas.

(i) O módulo deve permitir que o usuário escreva um comentário para cada regra avaliada.

(j) O módulo deve permitir que o usuário salve uma regra se a considerar interessante.

3. **Requisitos para cálculo de medidas subjetivas**

(a) O módulo deve permitir que um usuário que tenha realizado uma avaliação subjetiva solicite o cálculo das medidas subjetivas.

(b) O módulo deve calcular os valores das seguintes medidas subjetivas: conformidade, antecedente inesperado, conseqüente inesperado e antecedente e conseqüente inesperados. Essas medidas devem ser previamente cadastradas no RulEE.

(c) O módulo deve calcular e armazenar o valor das quatro medidas subjetivas para cada regra do conjunto foco, que não foi excluída e que não pertença ao

subconjunto de regras potencialmente interessantes, em relação a cada regra do subconjunto de regras potencialmente interessantes avaliada pelo usuário, exceto as regras avaliadas como conhecimento irrelevante.

4. **Requisitos para análise subjetiva das regras de um conjunto foco**

   (a) O módulo deve permitir que o usuário navegue pelas regras do conjunto foco que tiveram medidas calculadas e armazenadas.

   (b) Para que o usuário inicie a análise, o módulo deve apresentar as regras avaliadas divididas em dois grupos: regras classificadas como conhecimento óbvio/prévio e regras classificadas como conhecimento inesperado/útil. Junto com cada regra, o módulo também deve apresentar sua tradução (se existir), a avaliação feita pelo usuário, a indicação se foi salva como uma regra interessante e o comentário escrito no momento da avaliação, além de links para medidas subjetivas que foram calculadas com base nessa regra.

   (c) O módulo deve permitir que o usuário tenha acesso ao outro grupo de regras (óbvio/prévio ou inesperada/útil) do subconjunto de regras potencialmente interessantes avaliadas para dar continuidade à análise subjetiva.

   (d) O módulo deve permitir que o usuário altere o comentário de uma regra avaliada, salve-a como uma regra interessante ou marque-a como não interessante (caso tenha salvo como interessante anteriormente).

   (e) Para cada regra avaliada apresentada ao usuário, o módulo deve permitir que o usuário solicite ver as regras do conjunto foco que apresentem o valor de determinada medida subjetiva dentro de um intervalo indicado por ele.

   (f) O módulo deve apresentar ao usuário, quando solicitado, as regras do conjunto foco que, em relação a determinada regra avaliada, possuam valores dentro de um intervalo indicado pelo usuário para determinada medida. Juntamente com as regras devem ser apresentadas suas traduções (se existirem), os respectivos valores da medida subjetiva indicada, a informação se a regra foi salva como interessante e os comentários escritos sobre elas.

   (g) Para cada regra do conjunto foco apresentada, o módulo também deve permitir que o usuário salve-a como interessante e/ou escreva comentários sobre ela. O módulo também deve permitir editar essas informações.

   (h) Durante a análise subjetiva, o módulo deve permitir que o usuário visualize as regras que foram salvas como interessantes, durante a avaliação ou durante a análise, bem como os respectivos comentários.

5. **Requisitos para visualização de explorações realizadas**

(a) O módulo deve permitir que usuários de um mesmo projeto visualizem as explorações dos conjuntos de regras daquele projeto que já foram encerradas, desde que essas explorações estejam visíveis a todos os membros do projeto.

(b) O módulo deve permitir que um usuário visualize todas as explorações que ele criou e que já foram encerradas.

(c) Para cada exploração do projeto que pode ser visualizada, o módulo deve apresentar aos usuários o código e a descrição da exploração, bem como os objetivos e os interesses do usuário que realizou a exploração.

(d) O módulo deve permitir que o usuário veja a avaliação subjetiva feita pelo usuário responsável pela exploração e as regras que foram classificadas como interessantes, bem como os comentários escritos para cada uma delas.

6. **Requisitos para definição de traduções para um conjunto de regras**

(a) O módulo deve permitir que um usuário defina traduções para um conjunto de regras de um projeto que ele é membro.

(b) No momento da inserção de traduções, o módulo deve apresentar ao usuário os conjuntos de regras do projeto selecionado e solicitar que o usuário selecione um conjunto de regras.

(c) Caso já exista traduções inseridas para o conjunto de regras selecionado, o módulo deve informar o usuário.

(d) O módulo deve solicitar os seguintes textos para a definição de traduções: início (texto antes do antecedente), conectivo (texto de ligação entre itens), relação (texto entre antecedente e conseqüente) e fim (texto após o conseqüente). O módulo também deve questionar o usuário se os itens também devem ser traduzidos.

(e) O módulo deve armazenar para cada conjunto de regras os dados solicitados, além do código do usuário que inseriu a tradução.

(f) Caso o usuário informe que os itens também devem ser traduzidos, o módulo deve apresentar cada item individualmente e solicitar a sua tradução. Também deve dar a opção de não traduzir o item apresentado.

(g) O módulo deve armazenar as traduções dos itens fornecidas pelo usuário.

A.3.2 Requisitos de Desempenho

O requisitos de desempenho do módulo RulEE-SEAR são os mesmos requisitos determinados para o ambiente RulEE.

Apêndice

B

Casos de Uso do RulEE-SEAR

Nesse apêndice são apresentadas as especificações dos nove casos de uso definidos para o RulEE-SEAR. Esses casos de uso são apresentados na Figura B.1, na qual está indicado, para cada caso de uso, a tabela que apresenta sua especificação.

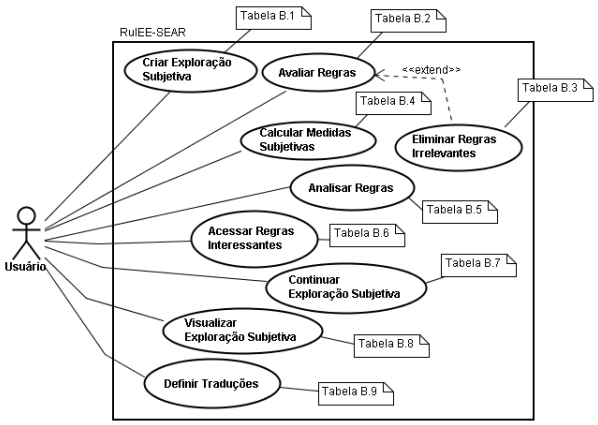


Figura B.1: Casos de uso do RulEE-SEAR

Na especificação de cada caso do uso, Criar Exploração Subjetiva (Tabela B.1), Avaliar Regras (Tabela B.2), Eliminar Regras Irrelevantes (Tabela B.3), Calcular Medidas Subjetivas (Tabela B.4), Analisar Regras (Tabela B.5), Acessar Regras Interessantes (Tabela B.6), Continuar Exploração Subjetiva (Tabela B.7), Visualizar Exploração Subjetiva (Tabela B.8) e Definir Traduções (Tabela B.9), são apresentados o propósito e a descrição do caso de uso, os requisitos, presentes no documento de requisitos apresentado no Apêndice A, que são satisfeitos pelo caso de uso e as ações do usuário, com as respectivas respostas do sistema, que podem ser encontradas no caso de uso em questão.

Caso de Uso:	**1**
Nome:	Criar Exploração Subjetiva
Ator:	Usuário
Propósito:	Criar uma nova exploração subjetiva para determinado conjunto de regras já inserido no RulEE.
Descrição:	O usuário solicita a criação de uma nova exploração subjetiva em um determinado projeto e indica os dados necessários. O sistema mostra os conjuntos de regras que podem ser utilizados e respectivos conjuntos foco e subconjunto de regras potencialmente interessantes. O usuário seleciona o conjunto que quer explorar, bem como o conjunto foco e o subconjunto de regras potencialmente interessantes. O sistema cria a exploração relacionando-a ao usuário. Ao final o usuário pode dar início à avaliação das regras (Caso de Uso 2 - Avaliar Regras)
Tipo:	Primário e essencial
Referências Cruzadas:	Requisitos: 1a, 1b, 1c, 1d Casos de Uso: 2
Fluxo Típico dos Eventos	
Ação do Ator	**Resposta do Sistema**
1. Este caso de uso inicia quando o usuário solicita a criação de uma nova exploração subjetiva a partir de um projeto do qual é membro.	2. O sistema apresenta os conjuntos de regras que podem ser utilizados, juntamente com os conjuntos foco e subconjuntos de regras potencialmente interessantes disponíveis.
3. O usuário seleciona um conjunto de regras, um conjunto foco e um subconjunto de regras potencialmente interessante.	4. O sistema solicita os dados que devem ser armazenados sobre a exploração que está sendo criada
5. O usuário fornece os dados solicitados.	6. Os dados são armazenados e o sistema apresenta uma mensagem indicando que a exploração foi criada com sucesso e pergunta se o usuário quer iniciar a avaliação das regras.
7. O usuário decide iniciar a avaliação das regras.	8. É ativado o caso de uso 2.
Fluxo Alternativo	
- Linhas 3 e 5: O usuário cancela a criação da exploração subjetiva, nada é armazenado e o sistema apresenta uma mensagem informando que a criação foi cancelada.	
- Linha 7: Usuário decide não iniciar a avaliação subjetiva das regras e sai do sistema.	

Tabela B.1: Especificação do caso de uso 1 (Criar Exploração Subjetiva)

Caso de Uso:	**2**
Nome:	Avaliar Regras
Ator:	Usuário
Propósito:	Realizar uma avaliação subjetiva das regras do subconjunto de regras potencialmente interessantes por meio de um questionário.
Descrição:	O usuário responde um questionário avaliando cada regra do subconjunto de regras potencialmente interessantes. O usuário também pode escrever comentários sobre qualquer regra e salvar aquelas que julgar interessante. Quando o usuário classifica uma regra como conhecimento irrelevante, é ativado o caso de uso 3 (Eliminar Regras Irrelevantes). Ao final o usuário pode solicitar o cálculo das medidas subjetivas (Caso de Uso 4 - Calcular Medidas Subjetivas).
Tipo:	Primário e essencial
Referências Cruzadas:	Requisitos: 1g, 2.a, 2b, 2c, 2d, 2i, 2j Casos de Uso: 1, 3, 4, 7

Fluxo Típico dos Eventos	
Ação do Ator	**Resposta do Sistema**
1. Este caso de uso tem início a partir do caso de uso 1 (Criar Exploração Subjetiva), do caso de uso 3 (Eliminar Regras Irrelevantes) ou do caso de uso 7 (Continuar Exploração Subjetiva).	2. O sistema ordena as regras do subconjunto de regras potencialmente interessantes pelo tamanho do *itemset* e apresenta a menor regra ao usuário, junto com dados necessários para a avaliação.
3. O usuário indica uma ou mais opções de avaliação, salva se for interessante e escreve um comentário se desejar.	4. O sistema salva os dados e apresenta a próxima regra ao usuário para ser avaliada (Linha 3).
5. Depois de avaliar todas as regras, o usuário solicita o cálculo das medidas subjetivas.	6. É ativado o caso de uso 4.

Fluxo Alternativo
- Linha 4: Se o usuário classificou a regra como conhecimento irrelevante é ativado o caso de uso 3.
- Linha 5: O usuário deseja salvar a avaliação e sair do sistema antes de calcular as medidas subjetivas.
- Linhas 3 e 5: O usuário decide encerrar a exploração subjetiva, o sistema salva os dados relacionados a essa exploração e a marca como encerrada.

Tabela B.2: Especificação do caso de uso 2 (Avaliar Regras)

Caso de Uso:	**3**
Nome:	Eliminar Regras Irrelevantes
Ator:	Usuário
Propósito:	Eliminar do conjunto foco as regras consideradas irrelevantes.
Descrição:	Se durante a avaliação das regras, o usuário classifica uma regra como conhecimento irrelevante, o sistema elimina do conjunto foco a regra irrelevante e as regras que possuem os itens que, segundo o usuário, tornam a regra irrelevante nas mesmas posições (antecedente ou conseqüente) da regra avaliada. O sistema também verifica se existem outras regras no conjunto foco que possuem esses itens em posições distintas e questiona o usuário se essas regras devem ser eliminadas. Ao final, o usuário pode continuar a avaliação das regras (Caso de Uso 2 - Avaliar Regras).
Tipo:	Primário e essencial
Referências Cruzadas:	Requisitos: 2e, 2f, 2g, 2h Casos de Uso: 2

Fluxo Típico dos Eventos	
Ação do Ator	**Resposta do Sistema**
1. Este caso de uso tem início a partir do caso de uso 2 (Avaliar Regras).	2. O sistema solicita do usuário os itens que tornam a regra irrelevante (se a regra possui mais de dois itens).
3. O usuário indica os itens que tornam aquela regra irrelevante.	4. O sistema elimina as regras do conjunto foco (incluindo as regras do subconjunto de regras potencialmente interessantes) que apresentam esses itens nas mesmas posições da regra avaliada, verifica se existe alguma regra no conjunto foco com esses itens em outras posições e questiona o usuário se ela também deve ser eliminada.
5. O usuário avalia a nova regra também com irrelevante.	6. O sistema volta ao passo 4 deste caso de uso.

Fluxo Alternativo
- Linha 4: O sistema não encontra outra regra com os itens em posições diferentes e é ativado o caso de uso 2.
- Linha 5: O usuário avalia a nova regra como não irrelevante, o sistema verifica que não há mais regras com esses itens e é ativado o caso de uso 2.

Tabela B.3: Especificação do caso de uso 3 (Eliminar Regras Irrelevantes)

Caso de Uso:	**4**
Nome:	Calcular Medidas Subjetivas
Ator:	Usuário
Propósito:	Calcular as medidas subjetivas em relação à avaliação subjetiva das regras do subconjunto de regras potencialmente interessantes realizada pelo usuário.
Descrição:	Os valores das medidas subjetivas são calculados para as regras do conjunto foco que não foram eliminadas e que não pertencem ao subconjunto de regras potencialmente interessantes, com base nas regras avaliadas (exceto aquelas avaliadas como conhecimento irrelevante). Após o cálculo o usuário pode iniciar a análise subjetiva das regras do conjunto foco (Caso de Uso 5 - Analisar Regras).
Tipo:	Primário e essencial
Referências Cruzadas:	Requisitos: 1g, 3a, 3b, 3c Casos de Uso: 2, 5, 7

Fluxo Típico dos Eventos	
Ação do Ator	**Resposta do Sistema**
1. Este caso de uso tem início a partir do caso de uso 2 (Avaliar Regras) ou do caso de uso 7 (Continuar Exploração Subjetiva).	2. O sistema calcula e armazena os valores das medidas subjetivas para cada regra do conjunto foco que não tenha sido eliminada. O sistema apresenta uma mensagem indicando que o processamento foi concluído com sucesso e pergunta se o usuário deseja iniciar a análise subjetiva das regras.
3. O usuário decide iniciar a análise das regras.	4. É ativado o caso de uso 5.

Fluxo Alternativo
- Linha 3: O usuário decide não iniciar a análise e sai do sistema.
- Linha 3: O usuário decide encerrar a exploração subjetiva, o sistema salva os dados relacionados a essa exploração e a marca como encerrada.

Tabela B.4: Especificação do caso de uso 4 (Calcular Medidas Subjetivas)

Caso de Uso:	**5**
Nome:	Analisar Regras
Ator:	Usuário
Propósito:	Analisar as regras do conjunto foco com base na avaliação do subconjunto potencialmente interessante, realizada pelo especialista.
Descrição:	O usuário navega pelas regras do conjunto foco a partir de regras do subconjunto potencialmente interessantes já avaliadas, salvando aquelas que forem interessantes e escrevendo comentários. A partir de uma regra avaliada, o usuário seleciona a medida que deseja utilizar e o intervalo de valores que lhe interessa, então o sistema apresenta as regras do conjunto foco que, em relação àquela regra avaliada possui os valores da medida selecionada no intervalo definido. A qualquer momento durante a análise o usuário pode acessar as regras que já foram salvas como interessantes (Caso de Uso 6 - Acessar Regras Interessantes).
Tipo:	Primário e essencial
Referências Cruzadas:	Requisitos: 1g, 4a, 4b, 4c, 4d, 4e, 4f, 4g, 4h Casos de Uso: 4, 6, 7

Fluxo Típico dos Eventos	
Ação do Ator	**Resposta do Sistema**
1. Este caso de uso tem início a partir do caso de uso 4 (Calcular Medidas Subjetivas) ou do caso de uso 7 (Continuar Exploração Subjetiva).	2. O sistema apresenta ao usuário as regras do conjunto foco dividas em dois grupos: regras avaliadas como inesperada/útil e regras avaliadas como conhecimento óbvio/prévio. Para cada regra o sistema apresenta os dados cabíveis para a análise.
3. O usuário seleciona uma das regras.	4. O sistema apresenta as medidas subjetivas disponíveis e solicita os valores mínimo e máximo para a medida que o usuário deseja analisar.
5. O usuário seleciona uma medida subjetiva e define o intervalo que as regras que deseja ver devem estar.	6. O sistema apresenta a regra avaliada selecionada, bem como suas informações, e as regras do conjunto foco que possuem valores da medida selecionada, em relação à regra avaliada, no intervalo definido. Também apresenta os dados necessários para a análise.
7. O usuário salva as regras que achar interessante e escreve comentários para as regras que desejar, ou editar essas informações. Depois de analisar as regras, o usuário solicita ver as regras salvas como interessantes.	8. É ativado o caso de uso 6.

Fluxo Alternativo
- Linha 3: O usuário decide encerrar a exploração subjetiva, o sistema salva os dados relacionados a essa exploração e a marca como encerrada.
- Linha 7: O usuário decide escrever ou alterar os comentários da regra já avaliada e marcá-la (ou desmarcá-la) como interessantes.
- Linha 7: O usuário deseja continuar a análise, voltando às regras avaliadas. O sistema volta ao passo 2 deste caso de uso.
- Linha 7: O usuário não deseja continuar a análise e sai do sistema.

Tabela B.5: Especificação do caso de uso 5 (Analisar Regras)

Caso de Uso:	**6**
Nome:	Acessar Regras Interessantes
Ator:	Usuário
Propósito:	Visualizar as regras que já foram salvas como interessantes.
Descrição:	O usuário visualiza as regras que ele salvou como interessantes durante a avaliação ou durante a análise. Caso esteja satisfeito com esse conjunto, pode encerrar a exploração; caso contrário, ele pode continuar a análise das regras (Caso de Uso 5 - Analisar Regras).
Tipo:	Primário e essencial
Referências Cruzadas:	Requisitos: 1g, 4h Casos de Uso: 5, 7
Fluxo Típico dos Eventos	
Ação do Ator	**Resposta do Sistema**
1. Este caso de uso inicia quando o usuário solicita o acesso às regras que foram salvas como interessantes a partir do caso de uso 5 (Analisar Regras) ou do caso de uso 7 (Continuar Exploração Subjetiva).	2. O sistema apresenta as regras que foram salvas como interessantes, bem como os comentários escritos para cada.
3. O usuário decide sair do sistema.	4. O sistema é finalizado.
Fluxo Alternativo	
- Linha 3: O usuário deseja continuar a análise das regras, caso o processa ainda não tenha sido encerrado, e é ativado o caso de uso 5.	
- Linha 3: O usuário decide encerrar o processo de exploração, caso este ainda não tenha sido encerrado, e o sistema marca-o como encerrado.	

Tabela B.6: Especificação do caso de uso 6 (Acessar Regras Interessantes)

Caso de Uso:	**7**
Nome:	Continuar Exploração Subjetiva
Ator:	Usuário
Propósito:	Realizar etapas de um processo de exploração subjetiva que já tenha sido iniciado anteriormente.
Descrição:	O usuário solicita continuar determinada exploração subjetiva que ele tenha iniciado anteriormente e que ainda não tenha sido encerrada. O sistema verifica as etapas que já foram concluídas e apresenta ao usuário a etapa que deve ser realizada.
Tipo:	Primário e essencial
Referências Cruzadas:	Requisitos: 1e, 1f Casos de Uso: 2, 4, 5, 6
Fluxo Típico dos Eventos	
Ação do Ator	**Resposta do Sistema**
1. Este caso de uso inicia quando o usuário solicita a realização de uma exploração subjetiva já iniciada por ele e ainda não encerrada, indicando qual é essa exploração e o projeto ao qual ela pertence.	2. O sistema verifica a situação da exploração e de acordo com as etapas que já foram realizadas é ativado o caso de uso 2, 4 ou 5.
Fluxo Alternativo	
- Linha 2: A exploração já foi encerrada, o sistema informa o usuário e pergunta se ele quer ver as regras salvas como interessantes durante o processo de exploração. Em caso afirmativo, é ativado o caso de uso 6.	

Tabela B.7: Especificação do caso de uso 7 (Continuar Exploração Subjetiva)

Caso de Uso:	**8**
Nome:	Visualizar Exploração Subjetiva
Ator:	Usuário
Propósito:	Acessar a avaliação e as regras interessantes de uma exploração subjetiva realizada por esse ou por outros usuários do mesmo projeto.
Descrição:	O usuário solicita ver uma exploração subjetiva realizada por ele ou uma realizada por outro usuário de um projeto ao qual ele pertence e que esteja visível. O sistema apresenta as regras do conjunto de regras potencialmente interessante com a avaliação realizada e as regras do conjunto foco que foram salvas como interessantes pelo usuário que realizou a exploração.
Tipo:	Primário e essencial
Referências Cruzadas:	Requisitos: 5a, 5b, 5c, 5d

Fluxo Típico dos Eventos	
Ação do Ator	**Resposta do Sistema**
1. Este caso de uso inicia quando o usuário solicita a visualização de uma exploração subjetiva de um o projeto ao qual ele pertence.	2. O sistema apresenta as explorações daquele projeto realizadas por este usuário e as explorações visíveis daquele projeto realizadas por outros usuários. Apresenta também os dados da exploração para que o usuário saiba qual é o objetivo da mesma.
3. O usuário seleciona uma exploração.	4. O sistema apresenta as regras do conjunto de regras potencialmente interessantes, juntamente com a avaliação feita pelo usuário responsável pela exploração e outros dados cabíveis. O sistema também apresenta as regras salvas como interessantes, com os comentários.
Fluxo Alternativo	

Tabela B.8: Especificação do caso de uso 8 (Visualizar Exploração Subjetiva)

Caso de Uso:	**9**
Nome:	Definir Traduções
Ator:	Usuário
Propósito:	Definir traduções para as regras de determinado conjunto de regras já inserido no RulEE.
Descrição:	O usuário solicita a inserção de traduções em um determinado projeto. O sistema mostra os conjuntos de regras que podem ser utilizados. O usuário seleciona aquele para o qual deseja definir as traduções e fornece os dados necessários. O sistema armazena as traduções das regras e dos itens, caso os itens também sejam traduzidos, relacionando-os ao usuário.
Tipo:	Secundário e essencial
Referências Cruzadas:	Requisitos: 6a, 6b, 6c, 6d, 6e, 6f, 6g

Fluxo Típico dos Eventos	
Ação do Ator	**Resposta do Sistema**
1. Este caso de uso inicia quando o usuário solicita a definição de traduções a partir de um projeto do qual é membro.	2. O sistema apresenta os conjuntos de regras pertencentes ao projeto selecionado.
3. O usuário seleciona o conjunto de regras para o qual deseja inserir as traduções.	4. O sistema solicita os dados que devem ser armazenados para traduzir as regras e questiona se os itens também devem ser traduzidos
5. O usuário fornece os dados solicitados e informa que deseja traduzir os itens.	6. Os dados são armazenados e o sistema apresenta os itens individualmente, solicitando a tradução
7. Para cada item apresentado, o usuário informa a tradução ou indica que o item não deve ser traduzido.	8. O sistema armazena as traduções dos itens e apresenta uma mensagem indicando que as traduções foram inseridas com sucesso.

Fluxo Alternativo
- Linha 4: O conjunto de regras selecionado pelo usuário já tem tradução definida, o sistema informa ao usuário e solicita a seleção de outro conjunto. O usuário seleciona um novo conjunto ou sai do sistema.
- Linha 5: O usuário informa que não deseja traduzir os itens, o sistema armazena os dados para tradução das regras e apresenta uma mensagem informando que a tradução foi inserida com sucesso.

Tabela B.9: Especificação do caso de uso 9 (Definir Traduções)