Pós-processamento de Regras de Associação usando Taxonomias

MARCOS AURÉLIO DOMINGUES[1]
SOLANGE OLIVEIRA REZENDE[2]

[1]LIACC-NIAAD – Universidade do Porto
Rua de Ceuta, 118, 6º Andar – 4050-190 Porto, Portugal
marcos@liacc.up.pt
http://www.liacc.up.pt

[2]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Av. Trabalhador São-Carlense, 400, Cx. Postal 668 – 13560-970 São Carlos, SP, Brazil
solange@icmc.usp.br
http://www.icmc.usp.br

Resumo. O processo de Mineração de Dados possibilita que seus usuários finais possam analisar, compreender e usar o conhecimento extraído em um sistema inteligente ou como apoio em processos de tomada de decisão. Entretanto, muitos dos algoritmos utilizados geram uma enorme quantidade de padrões, dificultando a análise. Esse problema ocorre em Regras de Associação, uma técnica de Mineração de Dados que procura identificar todos os padrões intrínsecos ao conjunto de dados. Uma abordagem que pode auxiliar a análise de Regras de Associação é o uso de taxonomias durante o pós-processamento. Neste artigo são propostos o algoritmo $\mathcal{GART}$, que usa taxonomias para generalizar Regras de Associação, e o módulo computacional *RulEE-GAR*, que possibilita a análise das regras generalizadas.

Palavras-Chave: Mineração de Dados, Pós-processamento, Regras de Associação, Taxonomias.

Post-processing of Association Rules using Taxonomies

Abstract. The Data Mining process enables that end users can analyse, understand and use the extracted knowledge in an intelligent system or to support decision processes. However, many algorithms used in the process find large quantities of patterns, complicating the analysis of the patterns. This fact occurs with Association Rules, a Data Mining technique that tries to identify intrinsic patterns in large data sets. A method that can help the analysis of the Association Rules is the use of taxonomies in the step of knowledge post-processing. In this paper, we propose the $\mathcal{GART}$ algorithm, which uses taxonomies to generalize Association Rules, and the *RulEE-GAR* computational module, that enables the analysis of the generalized rules.

Keywords: Data Mining, Post-processing, Association Rules, Taxonomies.

Received May 29, 2005 / Accepted July 21, 2005.

1 Introdução

A evolução da computação impulsionada pelo aumento do poder de processamento dos computadores, pelo armazenamento contínuo de grandes quantidades de dados a baixo custo, pela introdução de novas tecnologias de transmissão e disseminação de dados etc., tem dado às organizações a capacidade de armazenar informações detalhadas sobre cada transação que efetuam, gerando grandes Bases de Dados. As organizações reconhecem o valor das informações contidas em suas Bases de Dados e têm investido na aquisição e desenvolvimento de ferramentas de análise que produzam informações úteis.

Durante anos, métodos predominantemente manuais têm sido utilizados para transformar dados em conhecimento. Porém, o uso desses métodos tem se tor-

nado dispendioso (em termos financeiros e de tempo), subjetivo e inviável, quando aplicados a grandes Bases de Dados.

Devido aos problemas com os métodos manuais, tornou-se necessário o desenvolvimento de processos de análise automática, como o Processo de Extração de Conhecimento de Bases de Dados ou Mineração de Dados. Esse processo, de natureza iterativa e interativa, tem despontado por seu desempenho em diversos domínios, na extração de padrões válidos, novos, e potencialmente úteis dos dados [10].

Embora tenha se tornado necessária a utilização do processo de Mineração de Dados para extrair padrões a partir de dados, sua aplicação pode gerar uma elevada quantidade de conhecimento (padrões), muitos dos quais podem não ser importantes, relevantes ou interessantes para o usuário. Fornecer ao usuário uma grande quantidade de padrões não é produtivo pois, geralmente, ele procura poucos padrões que sejam interessantes. Portanto, é de vital importância o desenvolvimento de técnicas de apoio no sentido de fornecer aos usuários apenas os padrões mais interessantes [18, 12].

O problema de se gerar grandes quantidades de padrões recebe uma maior ênfase em Regras de Associação, uma das técnicas de Mineração de Dados que recentemente tem despertado grande interesse [3]. Na área acadêmica pesquisas vêm sendo desenvolvidas com essa técnica e as organizações têm utilizado os resultados no comércio, em contratos de seguro, na saúde, no geoprocessamento, na biologia molecular entre outras áreas [13, 5, 17].

Uma abordagem para solucionar o problema da grande quantidade de padrões extraídos pela técnica de Regras de Associação é o uso de taxonomias [20, 13, 1]. As taxonomias refletem uma visão coletiva ou individual de como os itens podem ser hierarquicamente classificados, podendo ser utilizadas para eliminar regras não interessantes e/ou redundantes [1].

Diante desse contexto, neste artigo são propostos o algoritmo $\mathcal{GART}$ e o módulo computacional *RulEE-GAR*. O algoritmo $\mathcal{GART}$ (*Generalization of Association Rules using Taxonomies* – Generalização de Regras de Associação usando Taxonomias) utiliza taxonomias para generalizar Regras de Associação. Já o módulo *RulEE-GAR*, além de facilitar o uso do algoritmo $\mathcal{GART}$ durante a identificação de taxonomias e generalização de regras, provê funcionalidades para analisar as Regras de Associação generalizadas.

Este artigo está organizado da seguinte maneira: na Seção 2 é descrito de modo geral o processo de Mineração de Dados, abordando principalmente as etapas do processo referentes ao Pré-processamento dos dados, Extração de Padrões e Pós-processamento do conhecimento extraído. A técnica de mineração de Regras de Associação é descrita na Seção 3. Na Seção 4 são apresentados aspectos gerais do uso de taxonomias em Regras de Associação. O algoritmo $\mathcal{GART}$ é descrito na Seção 5. Já na Seção 6 é apresentado o módulo computacional *RulEE-GAR*. Na Seção 7 são apresentados os resultados de alguns experimentos realizados com o algoritmo $\mathcal{GART}$. Por fim, na Seção 8 são apresentadas as conclusões sobre este artigo e trabalhos futuros.

2 Mineração de Dados

O processo de identificação de conhecimento em Bases de Dados é conhecido como Extração de Conhecimento de Bases de Dados ou Mineração de Dados, sendo geralmente referenciado na literatura como *Knowledge Discovery in Databases (KDD)*. O processo de Mineração de Dados é definido em [10] como:

> *Processo de identificação de padrões válidos, inovadores, potencialmente úteis e principalmente compreensíveis em conjuntos de dados.*

A divisão do processo de Mineração de Dados em três grandes etapas: Pré-Processamento, Extração de Padrões e Pós-Processamento, proposta em [4] e adotada em [16] é ilustrada na Figura 1.

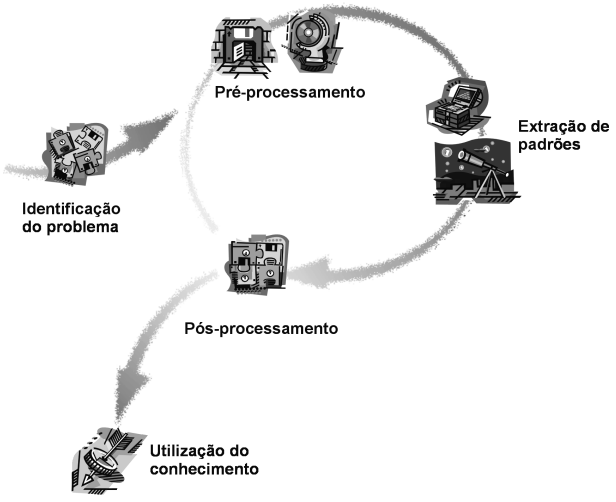

Figura 1: Etapas do processo de Mineração de Dados

A essa divisão são incluídas uma fase anterior ao processo de Mineração de Dados, referente à Identificação do Problema (nessa fase são definidos os objetivos a

serem alcançados e que tipo de informação se deseja extrair dos dados), e uma fase posterior ao processo, que se refere à Utilização do Conhecimento obtido, seja em um sistema inteligente ou como apoio em processos de tomada de decisão. As três etapas do processo de Mineração de Dados são apresentadas a seguir.

2.1 Pré-Processamento

Geralmente os dados selecionados para o processo de Mineração de Dados não estão em um formato adequado para a extração de conhecimento. Durante o processo de coleta de dados podem ocorrer diversos problemas que devem ser tratados, como erros de digitação, geração de dados incorretos ou inconsistentes por sensores, entre outros. Além desse fato, limitações de memória, tempo de processamento etc, podem impossibilitar a aplicação direta de alguns algoritmos de extração de padrões a todo o conjunto de dados. Todos esses problemas tornam necessário a utilização de métodos para tratamento, limpeza, redução do volume de dados, etc, antes de realizar a etapa de Extração de Padrões.

Na etapa de Pré-Processamento podem ser executadas diversas atividades no conjunto de dados:

Obtenção e unificação Consiste em unificar as diversas fontes de dados disponíveis para o processo de Mineração de Dados (arquivos-texto, planilhas, bancos de dados relacionais, *Data Warehouse* etc) em uma única fonte de dados;

Limpeza Consiste em aplicar técnicas de limpeza aos dados (por exemplo, a identificação e o tratamento de registros com valor inválido de algum atributo) com o objetivo de garantir a qualidade dos dados para o processo de Mineração de Dados;

Redução do volume Consiste em aplicar técnicas para reduzir o volume de dados a ser utilizado no processo de Mineração de Dados com o objetivo de viabilizar a utilização de algoritmos de extração de padrões que tenham restrições quanto ao espaço em memória, tempo de processamento etc.

2.2 Extração de Padrões

A etapa de Extração de Padrões é direcionada a cumprir os objetivos definidos na fase de Identificação do Problema. Nessa etapa são realizadas a escolha, a configuração e a execução de um ou mais algoritmos para a extração de conhecimento.

Por tratar-se de um processo iterativo, pode ser necessário que essa etapa seja realizada várias vezes para ajustar o seu conjunto de parâmetros visando a obtenção de resultados mais adequados aos objetivos pré-estabelecidos. Os ajustes podem, por exemplo, melhorar a precisão ou a compreensibilidade do conhecimento extraído. As atividades da etapa de Extração de Padrões são descritas a seguir:

Escolha da função Realizada de acordo com os objetivos desejáveis para a solução a ser encontrada. Há dois tipos de funções: preditivas e descritivas.

As funções preditivas consistem na generalização de exemplos com seus respectivos atributos meta conhecidos (atributo/classe a ser predita) em um modelo capaz de predizer o atributo meta de um novo exemplo. Já as funções descritivas consistem na identificação de padrões intrínsecos ao conjunto de dados, sendo que esses dados não possuem seus atributos meta especificados.

Escolha do algoritmo Realizada de acordo com a função de Mineração de Dados a ser utilizada e as especificações dos algoritmos disponíveis;

Obtenção de padrões Realizada aplicando-se o algoritmo escolhido para realizar a extração de padrões contidos no conjunto de dados.

2.3 Pós-Processamento

O Pós-Processamento é uma etapa importante do processo de Mineração de Dados na qual o conhecimento extraído pode ser simplificado, avaliado, visualizado ou simplesmente documentado para o usuário final.

Os métodos e procedimentos utilizados na etapa de Pós-Processamento são descritos a seguir:

Avaliação Consiste em avaliar o conhecimento extraído do conjunto de dados por meio de critérios, tais como precisão, compreensibilidade, interessabilidade, entre outros;

Interpretação e explanação Consiste em documentar, visualizar, modificar e/ou comparar (ao conhecimento pré-existente) o conhecimento extraído do conjunto de dados, de forma a torná-lo compreensível ao usuário;

Filtragem Consiste em filtrar o conhecimento que foi extraído do conjunto de dados, podendo ser realizada por vários mecanismos que variam de acordo com a técnica utilizada.

Após a análise do conhecimento na etapa de Pós-Processamento, o mesmo pode ser utilizado em um sistema inteligente ou como apoio em processos de tomada de decisão.

3 Regras de Associação

Na classificação usualmente empregada em Mineração de Dados, a técnica de Regras de Associação pode ser categorizada como uma *Função de Mineração de Dados Descritiva* [22, 16].

Uma Regra de Associação caracteriza o quanto a presença de um conjunto de itens nos registros de uma Base de Dados implica na presença de algum outro conjunto distinto de itens nos mesmos registros [2]. Desse modo, o objetivo das Regras de Associação é encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados. Por exemplo, observando os dados de vendas de um supermercado sabe-se que 80% dos clientes que compram o produto Q também adquirem, na mesma ocasião, o produto W. Nessa regra 80% corresponde a sua confiabilidade.

O formato de uma Regra de Associação pode ser representado como uma implicação $LHS \Rightarrow RHS$, em que LHS e RHS são, respectivamente, o lado esquerdo (*Left Hand Side*) e o lado direito (*Right Hand Side*) da regra, definidos por conjuntos disjuntos de itens. As Regras de Associação podem ser definidas como descrito a seguir [2]:

> *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 tal que $t_i \subseteq A$. A Regra de Associação é uma implicação na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$ e $LHS \cap RHS = \varnothing$. A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações T com confiança $conf$ se em $conf\%$ das transações de T em que ocorre LHS ocorre também RHS. A regra $LHS \Rightarrow RHS$ tem suporte sup se em $sup\%$ das transações em D ocorre $LHS \cup RHS$.*

O valor do suporte mede a força da associação entre LHS e RHS, e não relaciona possíveis dependências de RHS com LHS. Por outro lado, a confiança mede a força da implicação lógica descrita pela regra.

4 Uso de Taxonomias em Regras de Associação

O fato das Regras de Associação permitirem identificar associações entre itens e conjuntos de itens de uma Base de Dados faz com que os algoritmos produzam grandes quantidades de regras, muitas das quais não são interessantes para o usuário [13].

Devido a essa grande quantidade de regras, a análise e a compreensão do conhecimento torna-se difícil para o usuário. A aplicação de taxonomias em Regras de Associação pode ser utilizada para reduzir o volume de regras extraídas e, por conseqüência, facilitar a análise e compreensão do conhecimento.

As taxonomias refletem uma caracterização coletiva ou individual de como os itens podem ser hierarquicamente classificados [1]. Na Figura 2 é apresentado um exemplo de uma taxonomia. Nesse exemplo pode-se verificar que: camisetas são roupas leves, bermudas são roupas leves, roupas leves são um tipo de roupas, sandálias são um tipo de calçados, etc.

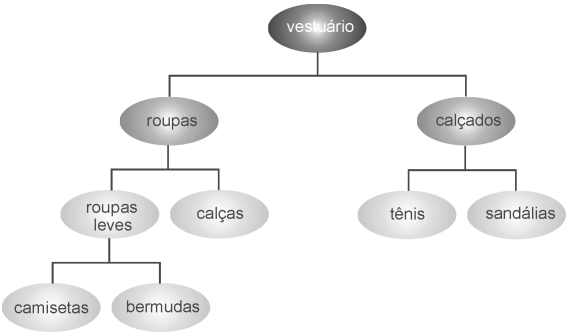

Figura 2: Exemplo de uma taxonomia para vestuário

Entretanto, existem algumas taxonomias cuja caracterização hierárquica dos itens é difícil de ser classificada. Um exemplo de uma taxonomia sem classificação é ilustrada na Figura 3, na qual pode-se verificar que é difícil identificar uma classificação para produtos1 e produtos2.

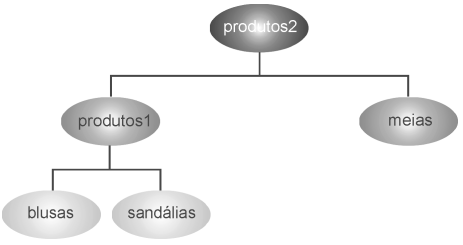

Figura 3: Exemplo de uma taxonomia sem classificação para vestuário

Entre as principais razões, apresentadas por [19], para o emprego de taxonomias em Regras de Associação, podem ser citadas:

- regras simples (cujos elementos são compostos apenas por itens terminais na taxonomia) podem não ter suporte suficiente para serem incluídas na solução, mas podem representar conhecimento inte-

ressante ao serem agrupadas segundo uma taxonomia;

- regras muito específicas podem ser generalizadas para melhorar a sua compreensibilidade;

- regras interessantes podem ser identificadas com o uso de informações contidas nas taxonomias. A interessabilidade de uma regra pode ser baseada em sua utilidade e inesperabilidade [18].

Uma Regra de Associação usando taxonomias pode ser definida como [20]:

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 tal que $t_i \subseteq A$. Seja $\mathcal{T}$ um grafo direcional e acíclico com os itens, representando um conjunto de taxonomias. Se há uma aresta em $\mathcal{T}$ de um item $a_p \in A$ para um item $a_c \in A$, a_p é dito ser ancestral de a_c e esse é dito ser descendente de a_p.
Uma Regra de Associação usando taxonomias é uma implicação na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$, $LHS \cap RHS = \varnothing$ e nenhum item em RHS é um ancestral de qualquer item em LHS. A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações T com confiança $conf$ se em $conf\%$ das transações de T em que ocorre LHS ocorre também RHS. A regra $LHS \Rightarrow RHS$ tem suporte sup se em $sup\%$ das transações de T ocorre $LHS \cup RHS$. É dito que uma transação t_i suporta um item $a_j \in A$, se a_j está em t_i ou a_j é um ancestral de algum item em t_i.

Na literatura há diversos algoritmos que podem ser utilizados para gerar Regras de Associação utilizando taxonomias: *Cumulate* e *Stratify* [20], *Genex* [21], *Prutax* [11] etc. Entretanto, os algoritmos apresentam o inconveniente de gerar todas as Regras de Associação possíveis (com e sem taxonomias) para, em seguida, utilizar uma medida subjetiva para remover as regras não interessantes. Como o valor da medida é definido pelo usuário, a mesma não garante que haverá uma redução do volume de regras. Buscando evitar esse inconveniente, na próxima seção é apresentado um algoritmo que foi proposto e implementado para utilizar taxonomias na generalização de Regras de Associação durante a etapa de Pós-processamento do conhecimento [9, 8, 6].

5 Algoritmo Proposto para Generalização de Regras de Associação

Analisando a estrutura das Regras de Associação, geradas por algoritmos que não utilizam taxonomias, é possível verificar que as mesmas podem ser generalizadas utilizando taxonomias. Esse fato é ilustrado na Figura 4.

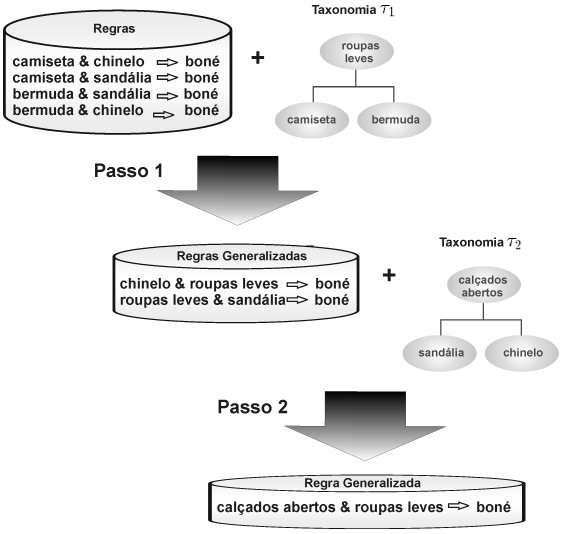

Figura 4: Generalização de Regras de Associação usando duas taxonomias

Inicialmente os itens camiseta e bermuda das regras

camiseta & chinelo ⇒ boné,
camiseta & sandália ⇒ boné,
bermuda & sandália ⇒ boné e
bermuda & chinelo ⇒ boné,

são substituídos pelo item roupas leves (que representa uma generalização) gerando duas regras chinelo & roupas leves ⇒ boné e duas regras roupas leves & sandália ⇒ boné. Em seguida, as regras repetidas são removidas permanecendo apenas as regras

chinelo & roupas leves ⇒ boné e
roupas leves & sandália ⇒ boné.

As duas regras resultantes do Passo 1 (Figura 4), são novamente generalizadas, sendo os itens chinelo e sandália substituídos pelo item calçados abertos (que representa outra generalização), gerando duas regras calçados abertos & roupas leves ⇒ boné. As regras repetidas são removidas permanecendo apenas uma Regra de Associação generalizada calçados abertos & roupas leves ⇒ boné

Devida a possibilidade de generalização de Regras de Associação, ilustrada na Figura 4, foi proposto um processo para generalizar as regras. O processo é apresentado na Figura 5.

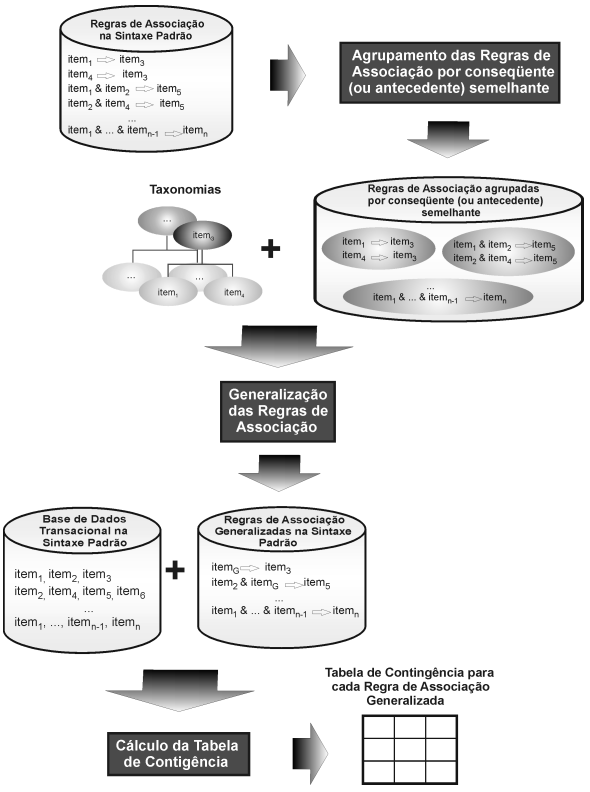

Figura 5: Processo proposto para generalização de Regras de Associação

O processo proposto generaliza apenas um dos lados (LHS ou RHS) das Regras de Associação. Inicialmente as regras (representadas na sintaxe padrão definida por [14]) são agrupadas em subconjuntos que apresentam antecedente ou conseqüente semelhante. Se o processo for utilizado para generalizar o lado esquerdo das regras (LHS), os subconjuntos são gerados utilizando conseqüentes (RHS) semelhantes e se o processo for utilizado para generalizar o lado direito das regras (RHS), os subconjuntos são gerados utilizando antecedentes (LHS) semelhantes. Na Figura 5 é ilustrado o processo de generalização do lado esquerdo das regras, por conseqüência, os subconjuntos são agrupados utilizando semelhanças no lado direito das regras. Em seguida são utilizadas as taxonomias para generalizar cada subconjunto, e, ao final da generalização, os itens de cada regra generalizada são ordenados lexicograficamente e a regra é armazenada em um conjunto de Regras de Associação generalizadas.

Ao final do processo, deve-se calcular a Tabela de Contingência para cada Regra de Associação que foi generalizada, uma vez que a sintaxe padrão proposta por [14] inclue a regra acompanhada de sua tabela. A Tabela de Contingência de uma regra representa a sua cobertura com relação a Base de Dados que foi utilizada na mineração da regra. Com o cálculo da Tabela de Contingência encerra-se o processo de generalização.

O algoritmo $\mathcal{GART}$ (*Generalization of Association Rules using Taxonomies* – Generalização de Regras de Associação usando Taxonomias) responsável pelo processo de generalização de Regras de Associação ilustrado na Figura 5 é apresentado no Algoritmo 1. Esse algoritmo recebe como entrada um conjunto de Regras de Associação R, um conjunto de Taxonomias $\mathcal{T}$, uma Base de Dados D e a definição do $lado$ da regra a ser generalizado – lado esquerdo (antecedente da regra) ou lado direito (conseqüente da regra). A notação $\overline{lado}$ define o lado das regras que não será generalizado.

Algoritmo 1 $\mathcal{GART}$

Requisitos: Um conjunto de Regras de Associação R, um conjunto de Taxonomias $\mathcal{T}$, uma Base de Dados D e a definição do $lado$ da regra a ser generalizado – lado esquerdo (antecedente das regras) ou lado direito (conseqüente das regras).

1: $Rg := \varnothing$;
2: $\widehat{E}$:= gera-subconjuntos(R, $\overline{lado}$);
3: **para todo** subconjunto $E \subseteq \widehat{E}$ **faça**
4: generaliza-regras(E, $\mathcal{T}$, $lado$);
5: ordena-lexicograficamente(E, $lado$);
6: $Rg := Rg \cup E$;
7: **fim-para**
8: **para todo** regra $r \in Rg$ **faça**
9: **se** r é uma regra generalizada **então**
10: calcula-TC(r, $\mathcal{T}$, D);
11: **fim-se**
12: **fim-para**
13: **retorna** Rg;

Inicialmente o algoritmo inicializa com "$\varnothing$" (vazio) o conjunto Rg que irá armazenar as Regras de Associação generalizadas. Em seguida são gerados os subconjuntos de regras que possuem antecedente ou conseqüente semelhante. Se o algoritmo for utilizado para generalizar o lado esquerdo das regras, os subconjuntos são gerados utilizando conseqüentes semelhantes e se o algoritmo for utilizado para generalizar o lado direito das regras, os subconjuntos são gerados utilizando antecedentes semelhantes. A geração dos subconjuntos é realizada com a função gera-subconjuntos (linha 2 do Algoritmo 1). Nessa função, o parâmetro $\overline{lado}$ indica que os subconjuntos são gerados utilizando o lado

das regras que não será generalizado. A função gera-subconjuntos é descrita em [7].

O próximo passo do algoritmo (linhas 3 a 7 do Algoritmo 1) consiste na generalização das Regras de Associação contidas em cada um dos subconjuntos de regras $E \subseteq \widehat{E}$, na ordenação lexicográfica dos itens de cada regra generalizada pertencente aos subconjuntos de regras e no armazenamento das regras no conjunto Rg. A generalização é realizada utilizando a função generaliza-regras, descrita em [7]. A ordenação lexicográfica dos itens das regras é realizada utilizando a função ordena-lexicograficamente, que também é descrita em [7].

Após a generalização é calculada a Tabela de Contingência para cada Regra de Associação generalizada contida no conjunto Rg (linhas 8 a 12 do Algoritmo 1). O cálculo da Tabela de Contingência é realizado utilizando a função calcula-TC, descrita em [7]. Por fim, o algoritmo (na linha 13) retorna o conjunto de Regras de Associação generalizadas Rg.

Na próxima seção é apresentado o módulo computacional *RulEE-GAR*, que utiliza o algoritmo $\mathcal{GART}$ para generalizar Regras de Associação e também fornece funcionalidades para analisar as regras generalizadas.

6 O Módulo Computacional *RulEE-GAR*

No módulo computacional *RulEE-GAR* [9, 8, 6], a generalização das Regras de Associação é realizada utilizando o algoritmo $\mathcal{GART}$, descrito na seção anterior. Já para prover as funcionalidades de análise das regras generalizadas, o módulo utiliza a Base de Dados e a Biblioteca de Classes desenvolvidas para o Ambiente de Exploração de Regras *RulEE* [15].

O ambiente *RulEE* foi desenvolvido com o objetivo de viabilizar tanto a análise quanto a disponibilização de regras de Classificação, Regressão e Associação. Nesse ambiente, o processo de análise e disponibilização das regras está baseado na idéia de se utilizar um Repositório de Regras e Medidas, que é a entidade principal do ambiente. No repositório devem ser armazenados os conjuntos de regras, os valores das medidas e as consultas realizadas pelos usuários. A contextualização do módulo *RulEE-GAR* no ambiente *RulEE*, bem como as suas funcionalidades são ilustradas na Figura 6.

Na Figura 6, os elementos que estão no retângulo pontilhado correspondem ao ambiente *RulEE*. Como pode ser verificado, o módulo computacional *RulEE-GAR* foi desenvolvido como um Módulo de Pós-Processamento desse ambiente.

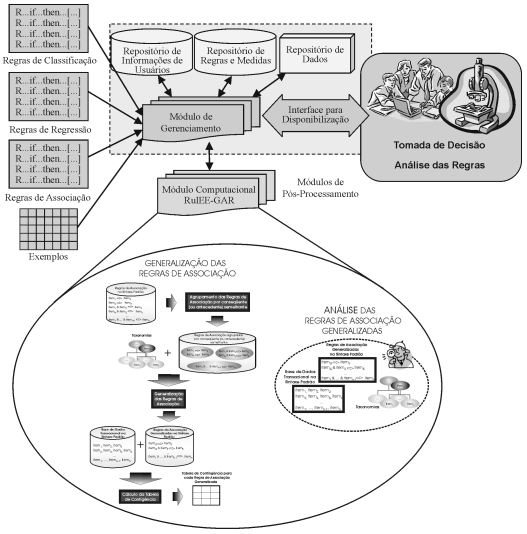

Figura 6: Contextualização e funcionalidades do módulo *RulEE-GAR* no ambiente *RulEE*

6.1 Descrição das Interfaces do Módulo *RulEE-GAR*

Na Figura 7 é ilustrada a interface do módulo *RulEE-GAR* que utiliza o algoritmo $\mathcal{GART}$ para generalizar os conjuntos de Regras de Associação armazenados no ambiente *RulEE*. A interface foi desenvolvida como um ferramenta *Wizard*, assim, o usuário inicialmente seleciona um conjunto de Regras de Associação a ser generalizado e, em seguida, define o conjunto de taxonomias que serão fornecidas ao algoritmo $\mathcal{GART}$. Por fim, o algoritmo $\mathcal{GART}$ é executado utilizando o conjunto de regras e de taxonomias definidos anteriormente.

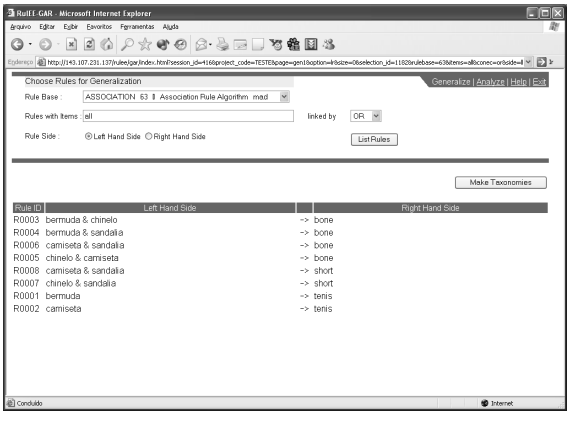

Figura 7: Interface para generalização de Regras de Associação

Já na Figura 8 é ilustrada a interface que permite o usuário analisar e explorar os conjuntos de Regras de Associação generalizadas.

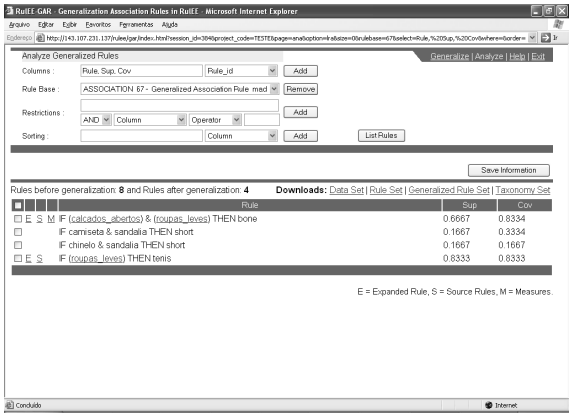

Figura 8: Interface para análise das Regras de Associação generalizadas

Na tela de análise de regras generalizadas (Figura 8) há alguns campos para entrada de dados que permitem o usuário selecionar um conjunto de regras generalizadas para ser analisado. Além de possibilitar o usuário selecionar um determinado conjunto de regras, a interface disponibiliza quatro *links* na seção *Downloads* para o usuário visualizar ou fazer o *download* dos arquivos contendo, respectivamente, o conjunto de dados transacionais (*Data Set*), o conjunto de regras originais (*Rule Set*), o conjunto de regras generalizadas (*Generalized Rule Set*) e o conjunto de taxonomias utilizadas (*Taxonomy Set*). O formato dos arquivos é descrito em [7].

Além dos *links* para visualização ou *download* dos arquivos, cada Regra de Associação generalizada listada pela consulta, apresenta alguns *links* que permitem o usuário explorar informações relacionadas a sua generalização. Os *links* que são descritos a seguir, estão posicionados a esquerda das regras (Figura 8):

Link Expanded Rule Representado na interface pela letra "E", permite o usuário visualizar a regra generalizada de modo expandida, ou seja, os itens generalizados em uma regra são substituídos pelos respectivos itens específicos.

Link Source Rules Representado na interface pela letra "S", permite o usuário visualizar as regras originais que geraram uma Regra de Associação generalizada.

Link Measures Representado pela letra "M", é visualizado na interface apenas se o usuário seleciona as medidas suporte (*Sup*) e/ou confiança (*Cov*) em sua consulta, e essas apresentam valor inferior ao suporte e/ou confiança mínimos definidos para o processo de mineração do conjunto de regras, antes desse ser generalizado.

Como o ambiente *RulEE* possibilita que o usuário ao inserir um conjunto de Regras de Associação, possa também inserir o valores de suporte e confiança mínimos utilizados na mineração desse conjunto, o *Link Measures* permite ao usuário comparar os valores de suporte e confiança mínimos de um conjunto de regras não generalizadas com os valores de suporte e confiança de uma regra generalizada gerada a partir desse conjunto. Essa análise é importante, pois embora o valor de suporte de uma Regra de Associação generalizada seja sempre igual ou maior do que o valor do suporte mínimo definido na mineração das regras não generalizadas, a mesma garantia não ocorre para a medida confiança [6].

Na Figura 8 também é possível verificar que os itens generalizados em uma regra (que estão delimitados por parênteses) são apresentados como *links*, para que o usuário possa visualizar os itens originais que foram generalizados e geraram esse item.

Por fim, o usuário também tem a opção de salvar em um arquivo no formato texto, as informações listadas na consulta que ele considerar importante.

Na próxima seção são apresentados alguns resultados obtidos com o uso do algoritmo $\mathcal{GART}$ na generalização de Regras de Associação.

7 Experimentos Realizados

Para mostrar que o uso de taxonomias na generalização de um conjunto de Regras de Associação pode reduzir o volume desse conjunto, foram realizados alguns experimentos utilizando o algoritmo $\mathcal{GART}$.

Os experimentos foram realizados utilizando uma base real que contém dados sobre as operações de vendas de um supermercado. No momento da realização dos experimentos, o volume de dados armazenados na base era referente as vendas realizadas no período de três meses. Assim, além de utilizar a base considerando as vendas dos três meses, foram realizadas quatro partições da mesma para a realização dos experimentos. As partições foram realizadas considerando 1 dia, 7 dias, 14 dias e 1 mês de vendas.

Na mineração das Regras de Associação foi utilizada a implementação do algoritmo *Apriori* realizada por Chistian Borgelt[1]. Para a mineração das regras foi utilizado um valor de suporte mínimo igual a 0.5, confiança mínima igual a 0.5 e número máximo de 5 itens por regra. Os conjuntos de regras obtidos são descritos a seguir:

[1]Disponível para *download* no *site* `http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html`.

- RuleSet_1dia – com as 32668 regras geradas a partir da partição de 1 dia;
- RuleSet_7dias – com as 19166 regras geradas a partir da partição de 7 dias;
- RuleSet_14dias – com as 16053 regras geradas a partir da partição de 14 dias;
- RuleSet_1mes – com as 21505 regras geradas a partir da partição de 1 mês;
- RuleSet_3meses – com as 19936 regras geradas a partir da base considerando 3 meses de vendas.

Para a realização dos experimentos foram construídos manualmente 13 conjuntos de taxonomias, a partir da análise da Base de Dados e dos 5 conjuntos de Regras de Associação gerados. Cada conjunto de taxonomias foi submetido ao algoritmo $\mathcal{GART}$ com cada um dos conjuntos de regras. Na Figura 9 é ilustrado um gráfico que mostra as taxas de redução dos conjuntos de regras ao serem generalizados por cada um dos 13 conjuntos de taxonomias. Nessa figura, os conjuntos de taxonomias foram denominados de "Tax" seguido de um número de identificação, como: Tax01.

Como pode ser verificado na Figura 9, os experimentos apresentam taxas de redução, do volume de um conjunto de Regras de Associação, variando entre 3,98% e 17,55%.

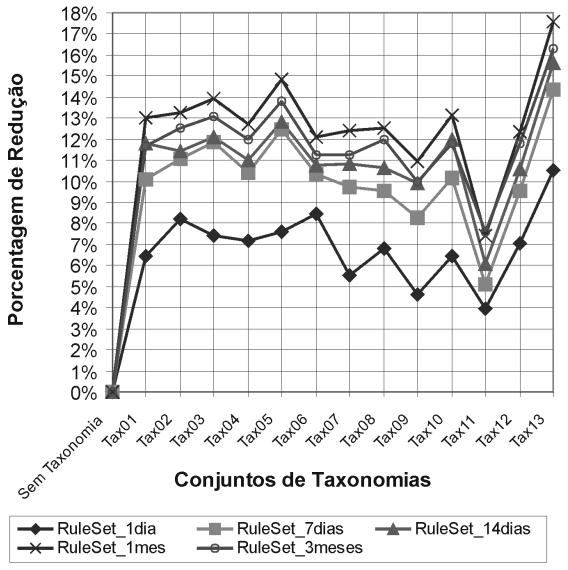

Figura 9: Taxas de redução obtidas com o uso de taxonomias na generalização das Regras de Associação

Nos conjuntos de Regras de Associação também foi possível verificar a presença de algumas taxonomias sem classificação. O uso do algoritmo $\mathcal{GART}$ com uma taxonomia sem classificação poderia resultar, por exemplo, em Regras de Associação generalizadas do tipo: calças & produtos1 $\Rightarrow$ camisetas.

A regra, do modo como é apresentada, não possui significado completo para o usuário, uma vez que é difícil identificar quais produtos fazem parte de produtos1. Entretanto, a regra pode apresentar significado completo, se a mesma for analisada utilizando a interface de análise de regras generalizadas do módulo computacional *RulEE-GAR* (descrito na Seção 6.1). Uma das funcionalidades dessa interface é permitir que o usuário possa verificar os itens originais que deram origem a um item generalizado, desse modo, essa funcionalidade pode ser utilizada para verificar os itens que estão contidos em produtos1, possibilitando, assim, a compreensão da regra.

Nesse contexto foram realizados alguns experimentos utilizando 3 conjuntos de taxonomias sem classificação. Na Figura 10 é ilustrado um gráfico que mostra as taxas de redução dos conjuntos de regras ao serem generalizados por cada um dos 3 conjuntos de taxonomias sem classificação, sendo que, cada conjunto foi denominado de "Tsc" seguido de um número de identificação, como: Tsc01.

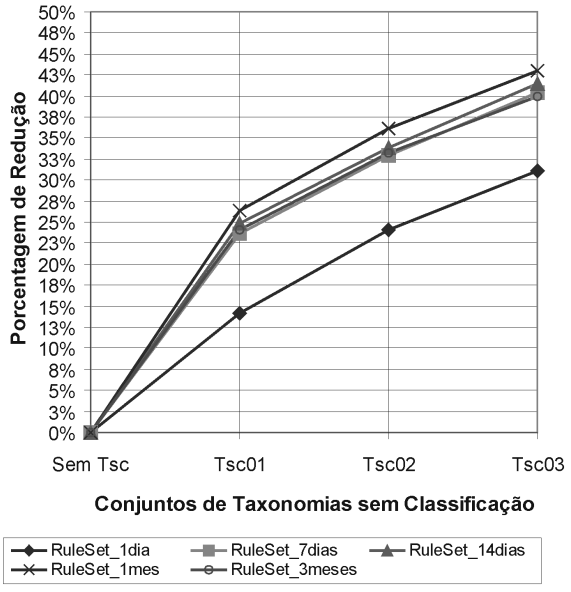

Figura 10: Taxas de redução obtidas com o uso de taxonomias sem classificação na generalização das Regras de Associação

Nos experimentos realizados, utilizando conjuntos de taxonomias sem classificação, as taxas de redução de volume dos conjuntos de Regras de Associação apre-

sentaram variação entre 14,12% e 42,97%.

Cabe ressaltar que como os conjuntos de taxonomias com e sem classificação são construídos pelo usuário, outros conjuntos podem gerar taxas de redução maiores do que as apresentadas nos experimentos, principalmente se os mesmos forem construídos por especialistas no domínio.

Em seguida foram realizados experimentos combinando alguns dos 13 conjuntos de taxonomias (que apresentam uma classificação) com os 3 conjuntos de taxonomias sem classificação. Para a realização dos experimentos foram utilizados os 3 conjuntos de taxonomias sem classificação e 6 conjuntos de taxonomias com classificação, resultando em 18 combinações. Das taxonomias que apresentam uma classificação foram escolhidos 2 conjuntos de taxonomias com o menor número de taxonomias, 2 conjuntos com um número mediano de taxonomias e 2 conjuntos com o maior número de taxonomias. Na Figura 11 é ilustrado um gráfico que mostra as taxas de redução dos conjuntos de regras ao serem generalizados utilizando combinações de conjuntos de taxonomias com e sem classificação. Cada combinação foi denominada de "C" seguida de um número de identificação, como: C01. As taxas de redução, apresentadas na Figura 11, variam entre 14,61% e 48,44%.

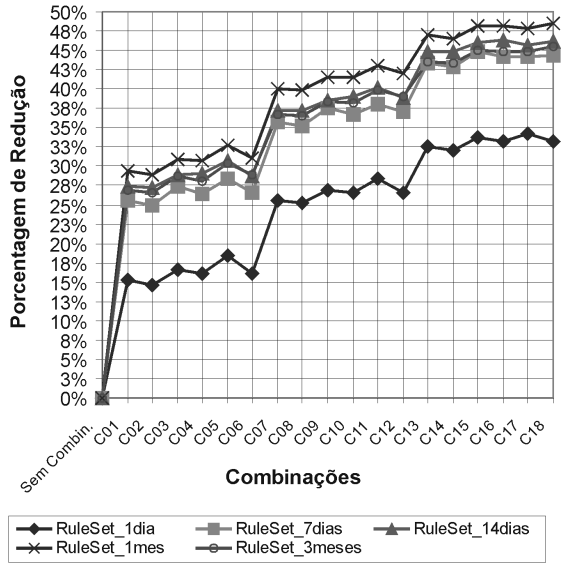

Figura 11: Taxas de redução obtidas com a combinação de taxonomias, com e sem classificação, na generalização das Regras de Associação

Por fim, foram realizados alguns experimentos utilizando a opção **-sort**, implementada neste trabalho para o algoritmo $\mathcal{GART}$. A opção **-sort** ordena (em ordem decrescente) o conjunto de taxonomias antes de generalizar o conjunto de Regras de Associação. A ordenação decrescente das taxonomias é realizada utilizando o número de especializações máximas (itens) das taxonomias como valores de comparação. Essa ordenação possibilita uma maior generalização das regras, uma vez que as taxonomias que apresentam os maiores números de especializações máximas (itens) serão consideradas primeiro durante o processo de generalização. Entretanto, para viabilizar o uso dessa opção, o conjunto de taxonomias precisa de duas ou mais taxonomias sobre o mesmo conjunto de itens. Por exemplo, considerando os itens short, bermuda e camiseta, o conjunto precisaria possuir taxonomias como:

roupa_esporte(short,bermuda) e
roupa_esporte(short,bermuda,camiseta).

Com relação aos conjuntos de taxonomias com e sem classificação, utilizados nos experimentos descritos, apenas Tax13 possui duas ou mais taxonomias sobre o mesmo conjunto de itens. Desse modo, foi realizado um experimento no qual o conjunto Tax13 foi submetido ao algoritmo $\mathcal{GART}$ (utilizando a opção **-sort**) junto com cada um dos 5 conjuntos de Regras de Associação. O resultado desse experimento apresentou uma taxa de redução 1% maior do que as apresentadas na Figura 9 para o conjunto Tax13. Além disso, como as combinações C05, C11 e C17 apresentam em sua composição o conjunto de taxonomias Tax13, foram realizados alguns experimentos com essas três combinações, utilizando a opção **-sort** do algoritmo $\mathcal{GART}$. Os resultados dos experimentos apresentaram taxas de redução 4% maiores do que as apresentadas na Figura 11 para as combinações C05, C11 e C17. Além disso, a combinação C17 ao generalizar o conjunto de regras RuleSet_1mes apresentou uma taxa de redução de 50,11% do volume de regras desse conjunto. Assim, pode-se verificar que a opção **-sort**, implementada para o algoritmo $\mathcal{GART}$, proporciona uma maior redução do volume de um conjunto de Regras de Associação.

8 Conclusões e Trabalhos Futuros

Um dos problemas encontrados no processo de Mineração de Dados é que muitos dos algoritmos utilizados geram uma enorme quantidade de padrões, dificultando consideravelmente sua análise. Esse problema recebe uma maior ênfase em Regras de Associação, uma vez que essa técnica de Mineração de Dados procura identificar todos os padrões intrínsecos ao conjunto de dados.

O uso de taxonomias no Pós-processamento das regras, para generalizar e eliminar regras não interessantes e/ou redundantes, pode auxiliar a análise das Regras de Associação. Neste artigo foram propostos o al-

goritmo $\mathcal{GART}$, que utiliza taxonomias para generalizar Regras de Associação, e o módulo computacional *RulEE-GAR*, que provê funcionalidades para analisar as regras generalizadas.

Nos experimentos realizados foi verificado que a generalização utilizando taxonomias com e sem classificação, bem como a combinação de ambas as taxonomias, possibilita a redução do volume dos conjuntos de Regras de Associação. Vale ressaltar que como os conjuntos de taxonomias com e sem classificação são construídos pelo usuário, outros conjuntos podem gerar taxas de redução maiores do que as apresentadas nos experimentos, principalmente se os mesmos forem construídos por especialistas no domínio. Por fim, foi constatado que a opção **-sort**, implementada neste trabalho para o algoritmo $\mathcal{GART}$, proporciona um aumento na taxa de redução do volume dos conjuntos de regras, uma vez que essa opção ordena (em ordem decrescente) o conjunto de taxonomias antes de generalizar o conjunto de Regras de Associação. A ordenação possibilita uma maior generalização das regras, pois as taxonomias que apresentam os maiores números de especializações máximas (itens) serão consideradas primeiro durante o processo de generalização.

A seguir são apresentadas algumas linhas de possíveis trabalhos futuros:

- Implementação de uma versão do algoritmo $\mathcal{GART}$ que permita a generalização de ambos os lados das Regras de Associação simultaneamente. A versão atual do algoritmo $\mathcal{GART}$ permite a generalização de apenas um lado das Regras de Associação (esquerdo ou direito);

- Incorporação de novos métodos de visualização de conjuntos de regras, ao módulo computacional *RulEE-GAR*, para facilitar a construção das taxonomias. Atualmente, o módulo *RulEE-GAR* permite visualizar conjuntos de regras ordenados por antecedentes ou conseqüentes semelhantes. Uma outra proposta seria o agrupamento das Regras de Associação utilizando algum algoritmo de *clustering* e a exibição das mesmas ordenadas por clusters;

- Desenvolvimento de um algoritmo para a construção automática ou semi-automática dos conjuntos de taxonomias para o algoritmo $\mathcal{GART}$. A contrução manual dos conjuntos de taxonomias consome um tempo considerável;

- Realização de experimentos utilizando outros conjuntos de dados para melhor verificar o uso de taxonomias na generalização e remoção de Regras de Associação não interessantes e/ou redundantes.

Agradecimentos

Os autores agradecem a Coordenação de Aperfeiçoamento de Nível Superior (CAPES) e a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo apoio financeiro.

Referências

[1] J. M. Adamo. *Data Mining for Association Rules and Sequential Patterns*. Springer-Verlag, New York, NY, 2001.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proceedings of Twentieth International Conference on Very Large Data Bases, VLDB*, pages 487–499, 1994.

[3] B. Baesens, S. Viaene, and J. Vanthienen. Post-processing of association rules. In *Proceedings of the Special Workshop on Post-Processing. The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2–8, 2000.

[4] J. A. Baranauskas. Extração automática de conhecimento por múltiplos indutores, 2001. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos, SP - Brasil.

[5] E. Clementini, P. Di Felice, and K. Koperski. Mining multiple-level spatial association rules for objects with a broad boundary. *Data & Knowledge Engineering*, 34(3):251–270, 2000.

[6] M. A. Domingues. Generalização de regras de associação, 2004. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos, SP - Brasil.

[7] M. A. Domingues and S. O. Rezende. Descrição de um algoritmo para generalização de regras de associação, 2004. Relatório Técnico do ICMC/USP - Número 228.

[8] M. A. Domingues and S. O. Rezende. Generalização de regras de associação usando taxonomias. In *Anais do I Workshop de Computação da Região Sul*, Florianópolis, Brasil, 2004.

[9] M. A. Domingues, S. O. Rezende, and M. F. Paula. Implementação de taxonomias para regras de associação em um ambiente de pós-processamento. In *Proceedings of IV Workshop on Advances & Trends in AI for Problem Solving (ATAI 2003)*, Chillán, Chile, 2003.

[10] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, volume 1, pages 1–30. American Association for Artificial Intelligence, Menlo Park, CA, 1996.

[11] J. Hipp, A. Myka, R. Wirth, and U. Güntzer. A new algorithm for faster mining of generalized association rules. In Jan M. Zytkow and Mohamed Quafafou, editors, *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD '98)*, pages 74–82, Nantes, France, 1998.

[12] A. Jorge, J. Poças, and P. Azevedo. A post processing environment for browsing large sets of association rules. In Marko Bohanec, Branko Kavsek, Nada Lavrac, and Dunja Mladenic, editors, *ECML/PKDD'02 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, pages 53–64, Helsinki, 2002.

[13] B. Liu, W. Hsu, S. Chen, and Y. Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems & their Applications*, 15(5):47–55, 2000.

[14] E. A. Melanda. Pós-processamento de regras de associação, 2004. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos, SP - Brasil.

[15] M. F. Paula. Ambiente para exploração de regras, 2003. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos, SP - Brasil.

[16] S. O. Rezende, J. B. Pugliesi, E. A. Melanda, and M. F. Paula. Mineração de dados. In S. O. Rezende, editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, volume 1, pages 307–335. Barueri, SP: Editora Manole, 2003.

[17] T. Semenova, M. Hegland, W. Graco, and G. Williams. Effectiveness of mining association rules for identifying trends in large health databases. In *Workshop on Integrating Data Mining and Knowledge Management. ICDM'01: The 2001 IEEE International Conference on Data Mining*, 2001. Avaliable in `http://cui.unige.ch/~hilario/icdm-01/DM-KM-Final/Semenova.pdf`. Access in 11/01/2005.

[18] A. Silberschatz and A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In Usama M. Fayyad and Ramasamy Uthurusamy, editors, *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pages 275–281, 1995.

[19] R. Srikant. Association rules: Past, present and future. ICCS 2001 International Workshop on Concept Lattice-based theory, methods and tools for Knowledge Discovery in Databases, 2001. Invited Talk. Disponível em: `http://www.almaden.ibm.com/cs/people/srikant/talks/assoc.pdf`. Acesso em: 19/11/2004.

[20] R. Srikant and R. Agrawal. Mining generalized association rules. *Future Generation Computer Systems*, 13(2–3):161–180, 1997.

[21] I. Weber. On pruning strategies for discovery of generalized and quantitative association rules. In In Liu Bing, Wynne Hsu, and Wang Ke, editors, *Proceedings Knowledge Discovery and Data Mining Workshop (Pricai'98)*, 1998. Disponível em: `http://citeseer.nj.nec.com/25197.html`. Acesso em: 10/12/2004.

[22] S. M. Weiss and N. Indurkhya. *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann, San Francisco, CA, 1998.