

Um Ambiente para Exploração de Regras de Associação Generalizadas

Marcos A. Domingues¹, Matheus G. B. Figueiredo², Solange O. Rezende³

¹Laboratório de Inteligência Artificial e Ciência de Computadores – Universidade do Porto
Rua de Ceuta, 118, 6º Andar – 4050-190 Porto, Portugal

²Universidade Federal de Lavras
Campus Universitário – 37200-000 Lavras, MG, Brasil

³Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Av. Trabalhador São-Carlense, 400, Cx. Postal 668 – 13560-970 São Carlos, SP, Brasil

marcos@liacc.up.pt, garcia.figueiredo@gmail.com, solange@icmc.usp.br

Abstract. *The Data Mining (DM) 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 (AR), a technique that tries to identify intrinsic patterns in large data sets. A method that can help the analysis of the rules is the generalization of the AR. In this paper, we propose an environment to generalize AR and to analyse the generalized rules.*

Resumo. *O processo de Mineração de Dados (MD) 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 (RA), uma técnica de MD que procura identificar todos os padrões intrínsecos ao conjunto de dados. Uma abordagem que pode auxiliar a análise das regras é a generalização de RA. Neste artigo é apresentado um ambiente proposto para a generalização de RA e análise das regras generalizadas.*

1. Introdução

O processo de Mineração de Dados tem como objetivo encontrar conhecimento, a partir de grandes conjuntos de dados, para ser utilizado em um sistema inteligente ou como apoio em processos de tomada de decisão [Fayyad et al. 1996, Rezende et al. 2003]. Assim, um requisito importante é que o conhecimento descoberto seja válido, além de compreensível, útil e interessante [Fayyad et al. 1996].

A aplicação do processo de Mineração de Dados para extrair conhecimento de dados, pode gerar uma elevada quantidade de 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. Buscando resolver esse problema, pesquisas em Pós-processamento

de conhecimento vêm sendo realizadas nos últimos anos [Bruha and Famili 2000, Baesens et al. 2000, Jorge et al. 2002, Domingues and Rezende 2005].

O problema relacionado com a geração de grandes quantidades de padrões recebe uma maior ênfase em Regras de Associação, uma das técnicas de Mineração de Dados que tem despertado grande interesse na área acadêmica e nas organizações. 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 [Agrawal and Srikant 1994]. Desse modo, o objetivo das regras é encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados. Uma Regra de Associação pode ser representada como uma implicação na forma $LHS \Rightarrow RHS$, em que LHS e RHS são, respectivamente, o lado esquerdo (*Left Hand Side*) e o lado direito (*Right Hand Side*) da regra, definidos por conjuntos disjuntos de itens.

Uma abordagem para resolver o problema da grande quantidade de padrões extraídos pela técnica de Regras de Associação é o uso de taxonomias [Srikant and Agrawal 1997, Liu et al. 2000, Adamo 2001]. As taxonomias podem ser utilizadas para eliminar regras não interessantes e/ou redundantes [Adamo 2001].

Diante desse contexto, neste artigo é proposto o ambiente \mathcal{ENGAR} (*Environment for Generalization and Analysis of Association Rules* - Ambiente para Generalização e Análise de Regras de Associação) que tem como objetivo fornecer funcionalidades para generalizar Regras de Associação e também para analisar as regras generalizadas. O artigo está organizado em 4 seções: na Seção 2. são apresentados alguns aspectos gerais do uso de taxonomias e a técnica de mineração de Regras de Associação generalizadas. O ambiente \mathcal{ENGAR} é descrito na Seção 3.. Por fim, na Seção 4. são apresentadas as conclusões sobre este artigo e propostos alguns trabalhos futuros.

2. Regras de Associação Generalizadas

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 [Liu et al. 2000].

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 consequência, facilitar a análise e compreensão do conhecimento.

As taxonomias refletem uma caracterização coletiva ou

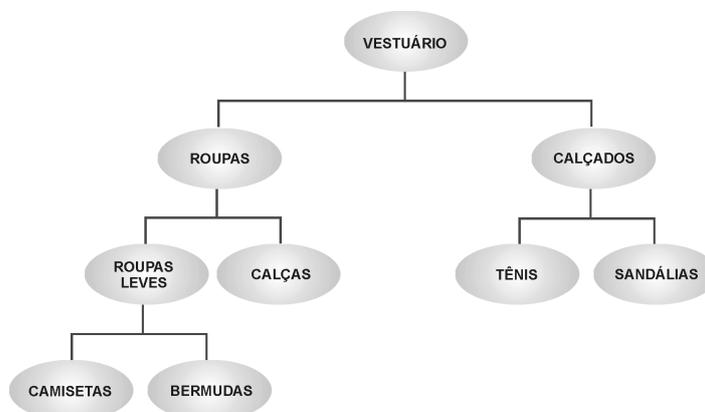


Figura 1. Exemplo de uma taxonomia para vestuário

individual de como os itens podem ser hierarquicamente classificados [Adamo 2001]. Na Figura 1 é apresentado um pequeno exemplo de uma taxonomia. Nesse exemplo pode-se verificar que: *camiseta* é uma *roupa leve*, *bermuda* é uma *roupa leve*, *roupa leve* é um tipo de *roupa*, *sandália* é um tipo de *calçado*, etc.

Entre as principais razões, apresentadas por [Srikant 2001], 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 interessante 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 [Silberschatz and Tuzhilin 1995].

Uma Regra de Associação usando taxonomias pode ser definida como [Srikant and Agrawal 1997]:

Seja D uma Base de Dados composta por um conjunto de itens $A = \{a_1, \dots, a_m\}$ ordenados lexicograficamente e por um conjunto de transações $T = \{t_1, \dots, t_n\}$, na qual cada transação $t_i \in T$ é composta por um conjunto de itens tal que $t_i \subseteq A$. 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 a_c é dito ser descendente de a_p .

Uma Regra de Associação usando taxonomias é uma implicação na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$, $LHS \cap RHS = \emptyset$ e nenhum item em RHS é um ancestral de qualquer item em LHS . A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações T com confiança $conf$ se em $conf\%$ das transações de T em que ocorre LHS ocorre também RHS . A regra $LHS \Rightarrow RHS$ tem suporte sup se em $sup\%$ das transações de T ocorre $LHS \cup RHS$. É 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* [Srikant and Agrawal 1997], *Prutax* [Hipp et al. 1998], *GART* [Domingues and Rezende 2005], etc.

3. O Ambiente \mathcal{ENGAR}

Nesta seção é apresentado o ambiente \mathcal{ENGAR} (*Environment for Generalization and Analysis of Association Rules* - Ambiente para Generalização e Análise de Regras de Associação) que tem como objetivo fornecer funcionalidades para generalizar Regras de Associação e também para analisar as regras generalizadas. A versão atual do sistema generaliza regras apenas por meio do algoritmo \mathcal{GART} [Domingues and Rezende 2005], que é apresentado no Algoritmo 1. Já a análise das regras generalizadas é realizada

por diversos métodos e medidas desenvolvidas a partir da tabela de contingência das regras [Lavrač et al. 1999].

Algoritmo 1 \mathcal{GART} . Fonte: [Domingues and Rezende 2005]

Require: 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 (consequente da regra).

```

1:  $Rg := \emptyset$ ; //A variável  $Rg$  irá armazenar o conjunto de regras generalizadas
2:  $\hat{E} :=$  gera-subconjuntos( $R, \overline{lado}$ ); //O parâmetro  $\overline{lado}$  indica o lado que não será generalizado
3: for all subconjunto  $E \subseteq \hat{E}$  do
4:   generaliza-regras( $E, \mathcal{T}, lado$ );
5:   ordena-lexicograficamente( $E, lado$ );
6:    $Rg := Rg \cup E$ ;
7: end for
8: for all regra  $r \in Rg$  do
9:   if  $r$  é uma regra generalizada then
10:    calcula-tabela-contingencia( $r, \mathcal{T}, D$ );
11:   end if
12: end for
13: return  $Rg$ ; //Retorna o conjunto de regras generalizadas

```

A tabela de contingência para uma regra, representa a sua cobertura com relação a Base de Dados que foi utilizada para minerar essa regra. Na Tabela 1, $n(LHS\ RHS)$ denota o número de transações no qual LHS e RHS são verdadeiros, $n(\overline{LHS}\ RHS)$ denota o número de transações no qual \overline{LHS} é falso e RHS é verdadeiro. O mesmo conceito se aplica para os demais elementos da tabela. N denota o número total de transações.

Tabela 1. Tabela de Contingência. Fonte: [Lavrač et al. 1999]

	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

3.1. Arquitetura

A arquitetura do ambiente \mathcal{ENGAR} é apresentada na Figura 2. O ambiente é basicamente composto por três módulos: entrada de dados, generalização e análise.

O módulo de entrada de dados é responsável por carregar no sistema os arquivos texto com os conjuntos de dados transacionais, Regras de Associação e taxonomias, a fim de que possam ser utilizados pelos demais módulos. O módulo de generalização é a parte do sistema responsável por executar os algoritmos de generalização de Regras de Associação. Por fim, o terceiro módulo contempla a funcionalidade de análise do sistema, onde um conjunto de regras previamente generalizado e armazenado no sistema pode ser avaliado através de vários métodos e medidas, como por exemplo, confiança, suporte, análise de correlação, dentre outras [Tan et al. 2005].

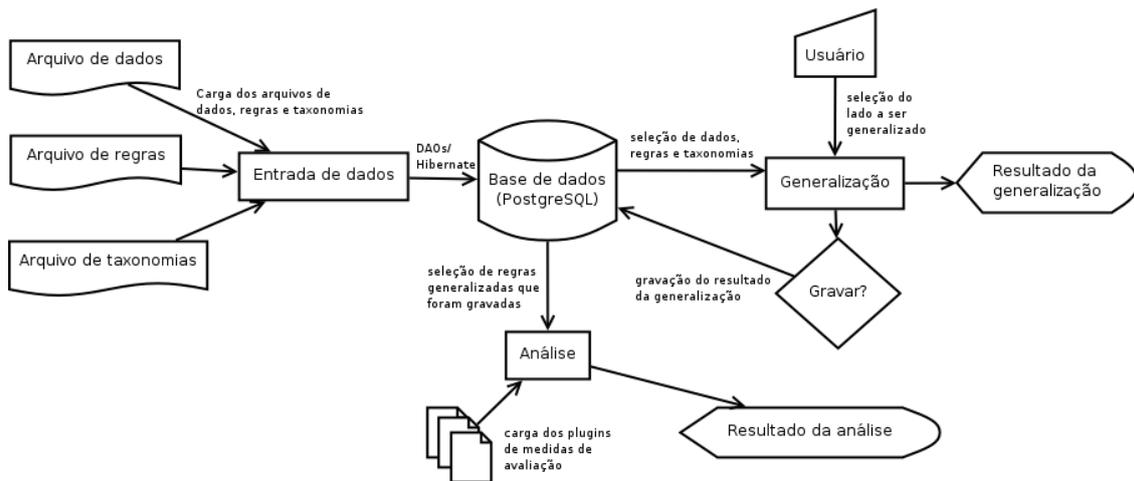


Figura 2. Arquitetura do ambiente *ENGAR*

3.2. Implementação

Na fase de implementação deste trabalho procurou-se seguir diretrizes de padrões de projeto e focar na reutilização de componentes já conhecidos e testados, desde que estes fossem distribuídos gratuitamente.

O sistema foi desenvolvido em Java, versão 1.5 (5.0). O desenvolvimento da interface gráfica foi feito através da biblioteca de componentes Swing, que já faz parte da distribuição padrão da linguagem Java. Para auxiliar no desenvolvimento e mesmo para acompanhar a execução do sistema foi utilizada a biblioteca Log4J (<http://logging.apache.org/log4j>), que permite a gravação de registros de log detalhados de acordo com o andamento do sistema.

O acesso ao banco de dados, em sua maior parte, foi feito através da ferramenta de mapeamento objeto-relacional Hibernate (<http://www.hibernate.org>). Uma das vantagens de sua utilização é liberar a programação de código SQL, e também facilitar a portabilidade do sistema entre vários Sistemas de Gerenciamento de Base de Dados (SGBD). Em sua versão atual a aplicação está disponível apenas para o SGBD PostgreSQL. O sistema também possui suporte à internacionalização, sendo que todas as mensagens exibidas pelo sistema (inclusive mensagens de log) são obtidas a partir de arquivos de propriedade do Java. Atualmente há definições apenas para Inglês e Português.

Para possibilitar a inclusão de novas medidas para avaliação de Regras de Associação generalizadas, o sistema implementa um mecanismo extensível para adição de novas medidas como plugins. Para isso, basta a criação e compilação de uma única classe em Java que implemente a medida no formato esperado pelo ambiente. É importante ressaltar que somente podem ser implementadas medidas cuja fórmula de cálculo se baseia nas informações da tabela de contingência das regras.

3.3. Interface

As funcionalidades do ambiente *ENGAR* estão distribuídas basicamente por três interfaces do sistema que são responsáveis pela entrada de dados, generalização das regras e sua posterior análise.

3.3.1. Entrada de Dados

A interface de entrada de dados permite ao usuário carregar arquivos de dados, de Regras de Associação e de taxonomias para a base de dados do sistema, tornando-os disponíveis para o restante do ambiente. Os arquivos deverão estar nos formatos padrões especificados em [Domingues and Rezende 2004] para serem processados pelo sistema.

O arquivo de dados contém o conjunto de dados transacionais que foram utilizados na mineração das Regras de Associação que serão generalizadas. Nesse arquivo cada transação é representada por uma linha do arquivo e os valores (itens) em cada transação são separados por um espaço em branco, como por exemplo:

```
camiseta chinelo bone tenis
camiseta sandalia bone tenis
bermuda camiseta
```

O arquivo de Regras de Associação, contém o conjunto de regras que será generalizado. O conjunto de regras está representado na sintaxe padrão para Regras de Associação definida em [Melanda and Rezende 2003]. Assim, cada linha do arquivo representa uma Regra de Associação que contém um identificador (*ID*), o lado esquerdo da regra (*LHS*), o lado direito da regra (*RHS*) e a frequência relativa dos valores $f(LHS\ RHS)$, $f(LHS\ \overline{RHS})$, $f(\overline{LHS}\ RHS)$ e $f(\overline{LHS}\ \overline{RHS})$ da tabela de contingência da regra [Lavrač et al. 1999], além do número total de transações consideradas *N*. Os valores da Tabela de Contingência são delimitados por “[]” (colchetes). Cada regra possui as informações apresentadas separadas por “;” (vírgula):

```
[R0001],camiseta,tenis,[0.500000,0.000000,0.000000,0.500000,6]
[R0002],bermuda & chinelo,bone,[0.166667,0.000000,0.333333,0.500000,6]
```

O arquivo de taxonomias contém o conjunto de taxonomias que será utilizado para generalizar as Regras de Associação. Cada linha do arquivo representa uma taxonomia que contém um item ancestral e, um ou mais itens descendentes na taxonomia. Os itens que representam descendentes são delimitados por “()” (parênteses) e separados por “;” (vírgula), como por exemplo:

```
roupas_sociais(calca,camisa,sapato)
calcados_sociais(sapato)
```

No momento da gravação dos arquivos de dados, o usuário deverá informar um “nome” para cada arquivo que estiver carregando, pois este nome servirá para identificá-lo unicamente dentro do ambiente, e opcionalmente uma descrição. Para o arquivo de Regras de Associação, poderá ainda informar quais foram os valores de confiança e suporte mínimos usados em sua mineração. Na Figura 3 é apresentada a interface de entrada de dados do ambiente *ENGAR*.

3.3.2. Generalização de Regras de Associação

Na Figura 4 é apresentada a interface de generalização de Regras de Associação do ambiente *ENGAR*. A interface de generalização, na versão atual do sistema, generaliza regras por meio do algoritmo *GART*. Para iniciar o processo, o usuário deverá selecionar o conjunto de Regras de Associação que deseja generalizar, o conjunto de taxonomias que serão

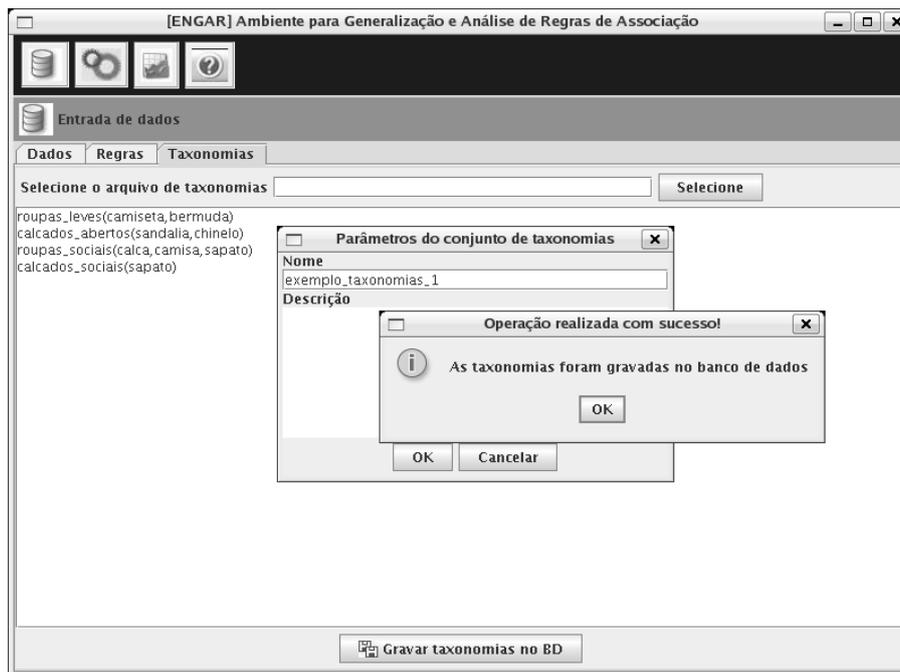


Figura 3. Interface para entrada de dados

usadas na generalização e o lado das regras na qual deseja que as taxonomias sejam aplicadas para que ocorra a generalização. A fim de poder calcular a tabela de contingência da regras generalizadas, o sistema também exige do usuário a seleção do conjunto de dados que foi utilizado na mineração das regras que serão generalizadas.

A seleção das regras, taxonomias e dados é feita através do “nome” informado na gravação dos respectivos arquivos na interface de entrada de dados. O conteúdo de cada arquivo selecionado poderá ainda ser visualizado pelo usuário antes de iniciar o processo de generalização das regras.

Assim que o processo de generalização for concluído o usuário poderá gravar o resultado na base de dados do sistema, permitindo sua posterior análise dentro do ambiente, ou poderá salvar o resultado em um arquivo texto, para ser novamente utilizado em outro processo de generalização.

3.3.3. Análise de Regras de Associação Generalizadas

Na Figura 5 é apresentada a interface de análise de regras generalizadas do ambiente *ENGAR*. A interface de análise das regras possibilita ao usuário avaliar o resultado de uma generalização previamente realizada. Estão disponíveis para seleção, cada um dos conjuntos de regras generalizados que foram gravados na interface de generalização, identificados pelo “nome” informado no momento de sua gravação.

O ambiente irá disponibilizar uma relação de medidas de avaliação disponíveis (confiança, suporte, análise de correlação, dentre outras), devendo o usuário selecionar aquelas que deseja em sua análise. Uma das características importantes do sistema é que o mesmo possui um mecanismo extensível para adição de novas medidas, que podem ser

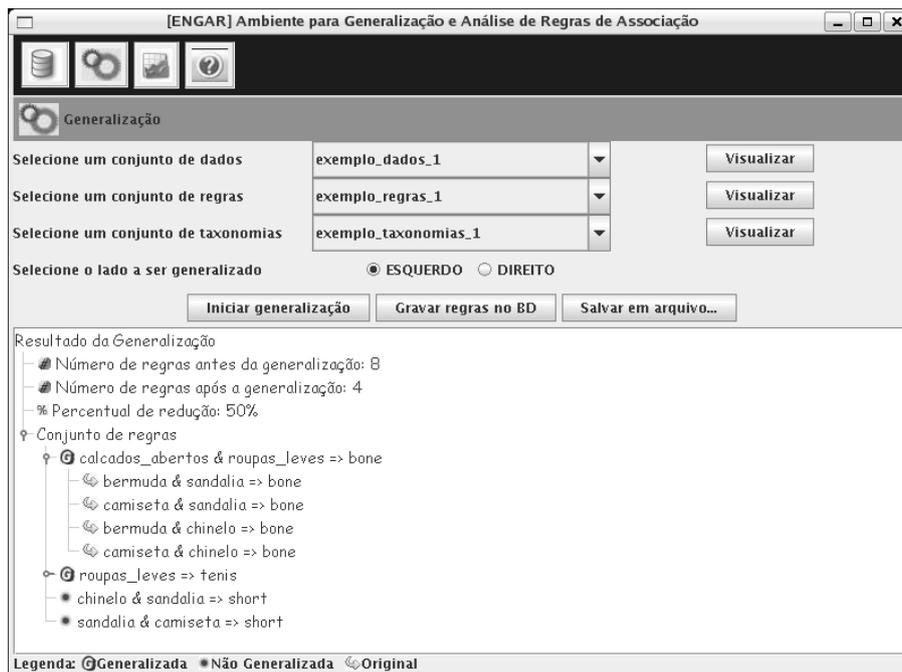


Figura 4. Interface para generalização de Regra de Associação

adicionadas como plugins. É importante ressaltar que somente podem ser implementadas medidas cuja fórmula de cálculo se baseia na tabela de contingência das regras.



Figura 5. Interface para análise de Regras de Associação generalizadas

O resultado da análise será exibido em uma tabela, onde cada linha representa uma regra generalizada e cada coluna representa o valor que essa regra apresenta para uma determinada medida selecionada. Além das medidas escolhidas pelo usuário, outras três

formas de avaliação estarão sempre presentes no resultado: a possibilidade de consultar a regra em formato expandido (substituindo os itens generalizados pelos itens que lhe deram origem), a consulta às regras originais e a comparação das medidas de suporte e confiança dessa regra generalizada em relação aos valores mínimos de suporte e confiança do conjunto de regras original que foi generalizado.

4. Conclusões e Trabalhos Futuros

No processo de Mineração de Dados, o uso de taxonomias no pós-processamento de Regras de Associação, para generalizar e eliminar regras não interessantes e/ou redundantes, pode auxiliar a análise das regras. Neste artigo foi apresentado um ambiente proposto para a generalização de Regras de Associação e análise das regras generalizadas.

Durante a realização deste trabalho se teve conhecimento de apenas um único sistema para análise de Regras de Associação generalizadas. O sistema foi proposto por Li Yang [Yang 2005], usa coordenadas paralelas para visualizar Regras de Associação simples e generalizadas, mas não apresenta funcionalidades para generalizar e analisar a qualidade das regras. O sistema aqui apresentado não possui recursos de visualização de regras generalizadas tão sofisticado quanto o apresentado por Li Yang [Yang 2005], mas fornece funcionalidades para generalizar, visualizar e analisar Regras de Associação; o que torna o sistema apresentado neste trabalho uma opção mais completa para o estudo de Regras de Associação generalizadas.

Como trabalhos futuros propõem-se a implementação de outras medidas de análise de Regras de Associação generalizada. Propõem-se também a implementação de um mecanismo extensível para adição de novos algoritmos de generalização de Regras de Associação, bem como a implementação de uma interface que permita a análise e comparação das generalizações realizadas pelos algoritmos (que poderão vir a ser implementados para o sistema). Por fim, pretende-se validar o sistema proposto através de experimentos usando conjuntos de dados reais e artificiais.

Agradecimentos: Ao POSC/EIA/58367/2004/Projeto Site-O-Matic (Fundação para Ciência e Tecnologia), co-financiado pelo FEDER.

Referências

- Adamo, J.-M. (2001). *Data Mining for Association Rules and Sequential Patterns*. Springer-Verlag, New York, NY.
- Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In Bocca, J. B., Jarke, M., and Zaniolo, C., editors, *Proceedings 20th International Conference on Very Large Data Bases, VLDB*, pages 487–499.
- Baesens, B., Viaene, S., and Vanthienen, J. (2000). Post-processing of association rules. In *Proceedings of the Special Workshop on Post-Processing. The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2–8.
- Bruha, I. and Famili, A. (2000). Postprocessing in machine learning and data mining. *ACM SIGKDD Explorations Newsletter*, 2(2):110–114.
- Domingues, M. A. and Rezende, S. O. (2004). Descrição de um algoritmo para generalização de regras de associação. Relatório Técnico do ICMC/USP - Número

228. Disponível em: ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/rel_tec/RT_228.pdf. Acesso em: 20/02/2007.
- Domingues, M. A. and Rezende, S. O. (2005). Post-processing of association rules using taxonomies. In *Proceedings of the Twelfth Portuguese Conference on Artificial Intelligence (EPIA-05)*, pages 192–197, Covilhã, Portugal. IEEE Computer Society.
- Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.
- Hipp, J., Myka, A., Wirth, R., and Güntzer, U. (1998). A new algorithm for faster mining of generalized association rules. In Zytkow, J. M. and Quafafou, M., editors, *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD '98)*, pages 74–82, Nantes, France.
- Jorge, A., Poças, J., and Azevedo, P. (2002). A post processing environment for browsing large sets of association rules. In Bohanec, M., Kavsek, B., Lavrac, N., and Mladenic, D., editors, *ECML/PKDD'02 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, pages 53–64, Helsinki.
- Lavrač, N., Flach, P., and Zupan, R. (1999). Rule evaluation measures: A unifying view. In Dzeroski, S. and Flach, P., editors, *Proceedings of the Ninth International Workshop on Inductive Logic Programming (ILP-99)*, pages 174–185. LNAI 1634.
- Liu, B., Hsu, W., Chen, S., and Ma, Y. (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems & their Applications*, 15(5):47–55.
- Melanda, E. A. and Rezende, S. O. (2003). Sintaxe padrão para representar regras de associação. Relatório Técnico do ICMC/USP - Número 206. Disponível em: ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/rel_tec/RT_206.pdf. Acesso em: 15/01/2007.
- Rezende, S. O., Pugliesi, J. B., Melanda, E. A., and Paula, M. F. (2003). Mineração de dados. In Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, volume 1, pages 307–335. Barueri, SP: Editora Manole.
- Silberschatz, A. and Tuzhilin, A. (1995). On subjective measures of interestingness in knowledge discovery. In Fayyad, U. M. and Uthurusamy, R., editors, *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pages 275–281.
- Srikant, R. (2001). Association rules: Past, present and future. ICCS 2001 International Workshop on Concept Lattice-based theory, methods and tools for Knowledge Discovery in Databases. Invited Talk. Disponível em: <http://www.rsrikant.com/talks/assoc.pdf>. Acesso em: 19/01/2007.
- Srikant, R. and Agrawal, R. (1997). Mining generalized association rules. *Future Generation Computer Systems*, 13(2–3):161–180.
- Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Yang, L. (2005). Pruning and visualizing generalized association rules in parallel coordinates. *IEEE Transactions on Knowledge and Data Engineering*, 17(1):60–70.