AMBIENTE DE PÓS-PROCESSAMENTO PARA REGRAS DE ASSOCIAÇÃO

JOÃO MANUEL POÇAS MARQUES DAS NEVES

Porto, Outubro de 2002

UNIVERSIDADE DO PORTO

FACULDADE DE ECONOMIA

AMBIENTE DE PÓS-PROCESSAMENTO PARA REGRAS DE ASSOCIAÇÃO

JOÃO MANUEL POÇAS MARQUES DAS NEVES

Sob orientação do Professor Doutor Alípio Jorge

Dissertação para a obtenção do grau de

Mestre em Análise de Dados e Sistemas de Apoio à Decisão

Porto, Outubro de 2002

À minha esposa Ana Margarida

Agradecimentos

A apenas algumas horas de concluir a presente dissertação, apraz-me relembrar algumas pessoas que, de uma forma ou de outra, contribuíram para a conclusão deste mestrado.

O Prof. Dr. Alípio Jorge, pelo seu apoio e pelos conselhos sensatos e oportunos.

Os amigos Rui Martins, Pedro Campos, Maria Manuel e Nuno Oliveira, pelos incansáveis incentivos.

Os meus pais que, uma vez mais, me deram, sem hesitar, toda a força para seguir em frente.

A Ana Margarida, amiga e esposa, cujo apoio e confiança constantes foram inexcedíveis.

Resumo

A obtenção de conhecimento, a partir das bases de dados de grandes dimensões existentes actualmente, resulta na produção de um elevado número de regras de associação, produzidas pelos algoritmos de descoberta de regras. Apesar de se considerarem de compreensão fácil e de aplicação em diversos domínios, tais como comércio, saúde, demografia, entre outros, a dificuldade de análise de um elevado número de regras de associação pode desencorajar os analistas a utilizar esta técnica.

No sentido de fomentar a utilização desta técnica nos mais diversos domínios, esta dissertação propõe uma metodologia e uma ferramenta de pós-processamento de regras de associação. A metodologia baseia-se na aplicação de um grupo de operadores que transformam conjuntos de regras em outros conjuntos de regras, permitindo, assim, direccionar a análise para uma determinada região do espaço de regras. A ferramenta proposta complementa a navegação pelo espaço de regras de associação com a representação gráfica de conjuntos de regras. Esta ferramenta foi implementada em ambiente Internet e permite analisar modelos de regras de associação representados no formato universal PMML.

Abstract

Due to the increasing size of mined databases, association rule engines produce a very large set of rules. Despite the fact that association rules are regarded as highly comprehensible and useful for data mining and decision support in fields such as marketing, retail, medicine, demographics, among others, lengthy outputs may discourage users from using the technique.

This thesis proposes a post-processing methodology and tool for browsing or visualizing large sets of association rules. This methodology is based on a set of operators that transform sets of rules into other sets of rules, allowing focusing on interesting regions of the rule space. The tool proposed allows each set of rules to be also analysed with different graphical representations. This is a web-based tool and can be used to analyse any association rules model presented in PMML format.

Índice

Índice das Tabelas

Índice das Figuras

Introdução

Actualmente, no decorrer das suas actividades quotidianas, empresas e organizações manipulam e armazenam uma grande quantidade de dados. Compreender esses dados, ou seja, conhecer a informação implícita nesses dados assume, cada vez mais, um papel de relevo no apoio à tomada de decisão.

A descoberta de regras de associação é uma técnica de *data mining*, que procura identificar determinados padrões de dados em bases de dados de grande dimensão, permitindo, após a sua interpretação, adquirir conhecimento específico acerca do problema em análise. Dada a dimensão das bases de dados actuais, o número de regras descobertas pode ser tão elevado, que quase transforma a sua interpretação num novo problema de *data mining*. Apesar dos estudos desenvolvidos por alguns investigadores neste domínio, ainda não foi encontrada uma solução óptima para a resolução deste problema.

Esta dissertação propõe uma nova metodologia e ferramenta de pós-processamento de regras de associação, cujo objectivo é atenuar o problema da interpretação de um elevado número de regras de associação. A metodologia tem por base a definição de um conjunto de operadores que permitem transformar determinados conjuntos de regras de associação em outros conjuntos de regras. A ferramenta que implementa esta metodologia (PEAR) permite, deste modo, direccionar a análise do utilizador de *data mining* para uma determinada região do espaço de regras, através da visualização de um conjunto de regras de associação de cada vez. A possibilidade de utilização de modelos de regras de associação representados em formato PMML, concede ao PEAR um carácter de universalidade. Esta ferramenta foi desenvolvida tendo como especial preocupação a disponibilização de um interface simples e intuitivo. Nesse sentido implementou-se uma aplicação que funciona em ambiente *web* que, cada vez mais, se assume como o meio privilegiado de comunicação e partilha de informação entre as pessoas.

Em termos de estrutura, a dissertação foi dividida em cinco capítulos, que serão resumidos seguidamente.

O primeiro capítulo começa por fazer uma abordagem a noções relacionadas com extracção de conhecimento e regras de associação. São analisados também alguns métodos de pós-processamento de regras de associação, que têm por principal objectivo facilitar a interpretação de um grande volume de regras de associação. Conclui-se este capítulo com uma referência à linguagem PMML, como forma de representação de diferentes modelos de *data mining* entre os quais as regras de associação.

No segundo capítulo, será introduzido o conceito de operadores de regras de associação. Esta metodologia de pós-processamento de regras baseia-se na utilização de um grupo de operadores que transformam um conjunto de regras num outro conjunto, permitindo focar a análise em determinadas regiões do espaço de regras. Será explicado o modo de funcionamento dos operadores, como podem ser aplicados e explicado o porquê da sua implementação num ambiente *web*.

O terceiro capítulo apresenta um sistema que implementa a metodologia apresentada no capítulo anterior: o PEAR. Serão discutidas algumas questões técnicas relacionadas com a implementação deste sistema: como funciona, que ambiente de implementação foi escolhido, que plataformas e ferramentas foram utilizadas e qual a sua estrutura funcional.

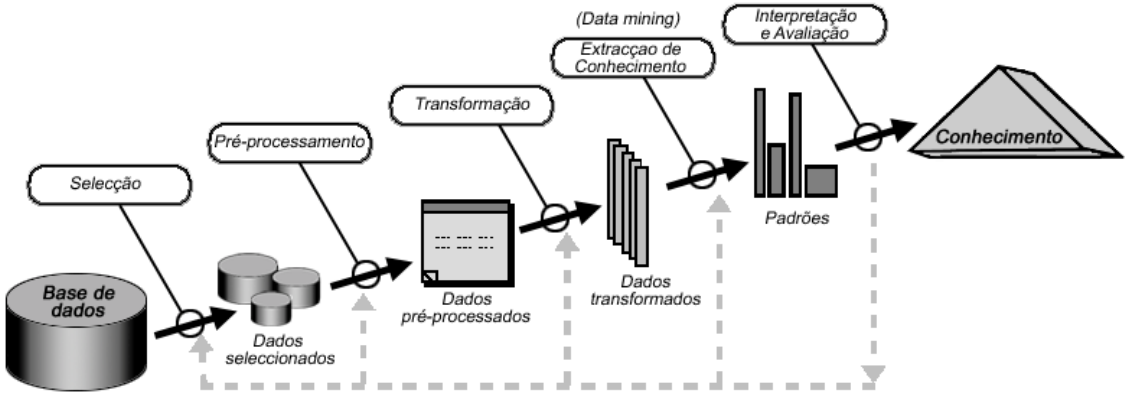
A metodologia proposta é analisada e avaliada no quarto capítulo. Inicia-se o capítulo por uma análise ao desempenho do sistema perante um elevado número de regras de associação. Utiliza-se, posteriormente, o PEAR perante um modelo real de regras de associação demonstrando-se uma possível navegação pelo espaço das regras. Apresentam-se, no final, alguns pontos fortes e fracos do PEAR, quer ao nível de desempenho quer ao nível de funcionalidade.

Finalmente retiram-se as necessárias ilações do trabalho desenvolvido, assim como se apresentam sugestões para possíveis desenvolvimentos futuros.

Capítulo I Extracção de conhecimento e regras de associação

O rápido desenvolvimento na área das tecnologias de informação a que se vem assistindo nos últimos anos tem contribuído de forma decisiva para o crescimento exponencial da quantidade de informação recolhida e armazenada em bases de dados (AGRAWAL, R. e R. SRIKANT, 1994). Actualmente, é fácil encontrar grandes empresas da área do comércio a executarem milhões de transacções por dia. Uma das maiores empresas norte-americanas, a *Wal-Mart*, por exemplo, executa mais de vinte milhões de transacções diárias (*Fonte: Wal-Mart Stores, Inc. at a Glance*). O problema da análise de quantidades de dados desta ordem de grandeza, por forma a servir de suporte à tomada de decisão, é o de encontrar métodos e algoritmos cujo desempenho se coadune com o aumento da dimensão das bases de dados.

O processo de descoberta de conhecimento (informação de interesse, desconhecida à partida, não trivial e potencialmente útil), a partir de dados disponíveis em bases de dados de grande dimensão, designa-se por *Knowledge Discovery in Databases* ou KDD (FAYYAD et al., 1996). Uma das etapas que fazem parte do processo de KDD, abaixo esquematizado, é a etapa de *data mining*, que consiste na extracção de informação, recorrendo a diversas técnicas conforme a situação em análise, de um conjunto de dados previamente processados.



Fonte: FAYYAD et al. 1996.

Figura 1 Esquema do processo de KDD, ilustrando as potenciais repetições e iterações.

Conforme ilustrado no esquema anterior (figura 1), o processo de extracção de conhecimento não é linear, uma vez que, em cada etapa, pode ser necessário refazer as etapas anteriores. As etapas que fazem parte do processo de extracção de conhecimento podem ser descritas, resumidamente, da seguinte forma:

a) Selecção

Esta etapa consiste na recolha e selecção dos dados necessários à análise do problema em questão. Na maior parte dos casos, trata-se de dados operacionais das empresas, o que pode dificultar a sua recolha. A eventual existência de bases de dados em diferentes estruturas e formatos, numa mesma organização constitui um outro obstáculo que, a este nível, se coloca.

b) Pré-processamento

Nesta fase, procura-se eliminar inconsistências (dados nulos ou repetidos) entre os dados seleccionados. O objectivo desta etapa é o de antecipar, à partida, a correcção de possíveis erros na análise dos dados.

c) Transformação

Após a etapa de pré-processamento, pode justificar-se a organização dos dados de uma forma harmonizada, preparando-os para a etapa seguinte. Nesta etapa, pode ser necessário incluir alguns dados externos à organização, tais como informação geográfica ou informação demográfica.

d) Extracção de conhecimento (*data mining*)

Esta é a fase em que realmente se desenrola a descoberta de conhecimento. Podem ser utilizadas diversas técnicas de *data mining*: regras de associação, *clustering*, redes neuronais, algoritmos genéticos, árvores de decisão e outras. A escolha da técnica a utilizar numa determinada análise depende das tarefas que se pretendem realizar (tarefas de classificação, tarefas de decisão, tarefas de comportamento), das características dos dados (quantidade de registos, quantidade de atributos), da

capacidade de estimar o significado estatístico do resultado e, também, dos recursos informáticos necessários para utilizar determinada técnica.

e) Interpretação e Avaliação

Uma vez produzidos os resultados na etapa anterior, pretende-se interpretá-los e retirar as respectivas conclusões. Nesta etapa, procura-se apresentar, de uma forma útil para os agentes decisores, os resultados da extracção de conhecimento.

A descoberta de regras de associação é uma das técnicas mais utilizadas na etapa de *data mining*. O estudo da descoberta de regras de associação foi introduzido por Agrawal, Imielinsky e Swami em Maio de 1993 (AGRAWAL et al., 1993). A sua aplicabilidade prática às diferentes áreas de negócio das organizações em conjunto com a fácil compreensão que lhe é inerente, até mesmo para não peritos em *data mining*, tem feito das regras de associação um método extremamente popular. O trabalho apresentado nesta dissertação situa-se na fase de Interpretação e Avaliação de modelos de *data mining*, também designada por pós-processamento. Em particular, os modelos considerados são conjuntos de regras de associação.

Neste capítulo, serão definidas as regras de associação, assim algumas das medidas que as caracterizam: suporte, confiança, *lift*, *coverage*, *leverage* e *conviction*. Será também abordado o principal problema relacionado com os algoritmos de obtenção de regras de associação (o da interpretação das regras produzidas), através da análise de diferentes estudos que se têm debruçado sobre este tema. Este capítulo debruça-se ainda sobre a linguagem PMML, que se trata de uma linguagem *standard* de representação de modelos preditivos de *data mining*, e de que forma pode ser utilizada na metodologia em análise nesta dissertação.

1 Descoberta de regras de associação

A descoberta de regras de associação permite encontrar padrões, associações ou correlações frequentes em conjuntos de itens (objectos) de uma base de dados transaccional, relacional ou de outros tipos de repositórios de informação. A motivação do estudo iniciado por AGRAWAL foi a necessidade de obtenção de conhecimento por parte de organizações da área do retalho (supermercados). Esta procura específica de conhecimento designa-se, em *data mining*, por *market basket analysis*. No entanto, a descoberta de regras de associação pode aplicar-se a diversas áreas de negócio tais como o estudo de dados dos recenseamentos da população, a análise de informação médica, o estudo dos acessos a computadores, entre muitos outros (MORZY, T. e M. ZAKRZEWICZ, 1997). A descoberta de regras de associação pode ser utilizada como suporte à tomada de decisão. Esta forma de descoberta de conhecimento é mais adequada a tarefas de *data mining* que não pretendam apenas satisfazer um único objectivo específico, pois permite que, diferentes agentes decisores possam obter diferentes perspectivas da mesma informação em análise. Através de uma mesma fonte de dados, um gestor pode optar por analisar o que caracteriza um bom cliente, determinar os produtos que certo tipo de clientes compram, identificar produtos que influenciem a venda de um determinado produto ou, simplesmente, caracterizar os seus grupos de clientes (BERRY e LINOFF, 2000). Obtém-se, frequentemente, informação adicional relevante, que não corresponde a nenhuma questão formulada à partida. As regras de associação procuram identificar uma relação entre objectos, designados por itens, passíveis de serem representados em bases de dados, e que possam, de alguma forma, estar relacionados entre si.

Um exemplo de regra de associação pode ser uma expressão do género: “De entre os utilizadores ao *site* do INE que consultaram estatísticas da saúde, cerca de 45% consultaram, na mesma sessão, estatísticas demográficas”. Esta afirmação pode ser representada de outro modo:

estatísticas da saúde -> estatísticas demográficas (confiança=0,45).

Na expressão anterior, as estatísticas da saúde e as estatísticas demográficas correspondem a itens, sendo que o item estatísticas da saúde (do lado esquerdo da implicação) designa-se por conjunto antecedente e o item estatísticas demográficas por conjunto consequente da regra. O valor 45% identifica a confiança desta regra, ou seja, a probabilidade de que, acontecendo uma consulta do conjunto antecedente (estatísticas da saúde) aconteça também uma consulta do consequente (estatísticas demográficas).

As regras de associação podem ser formalmente definidas da seguinte forma (adaptado de AGRAWAL, et al., 1993):

> Seja $I = \{i_1, i_2, \ldots, i_n\}$ um conjunto de objectos denominados itens que podem assumir valores binários 0 ou 1 (falso ou verdadeiro), conforme representem a presença ou não de um objecto em particular. Seja D um conjunto de transacções, em que cada transacção T corresponde a um conjunto de itens tal que $T \subseteq I$. Considera-se ainda que um conjunto de itens A está contido numa transacção T, se todos os itens do conjunto têm valor "verdadeiro" na transacção, ou seja, fazem parte dessa mesma transacção. Uma regra de associação R pode ser representada por uma expressão da forma: $A \rightarrow B$, onde $A \subseteq I$, $B \subseteq I$ e $A \cap B = \varnothing$. É ainda possível tratar as variáveis quantitativas ou qualitativas, criando intervalos de valores, utilizando-as, posteriormente, como binárias.

Um outro exemplo prático de uma regra de associação é a afirmação de que "70% das pessoas que compram polvo também compram molho de salsa". Uma regra deste género pode levar um determinado gerente de marketing de um supermercado a realizar, por exemplo, uma promoção para permitir escoar um dos produtos, vendendo em conjunto os dois artigos por um preço mais favorável. A facilidade de interpretação das regras de associação, aliada a uma utilidade prática muito forte, incentivou inúmeros investigadores a desenvolverem algoritmos de descoberta de regras de associação.

Os primeiros algoritmos a serem utilizados na descoberta de regras de associação foram o AIS (AGRAWAL et al., 1993) e SETM (HOUTSMA e SWAMI, 1993). No entanto,

depois desta data, novos algoritmos foram criados. O algoritmo padrão actualmente mais utilizado é sem dúvida o *Apriori* (AGRAWAL et al., 1994). Este algoritmo descobre todos os conjuntos de itens frequentes e produz todas as regras, dentro de certos valores de suporte e confiança, que possam ser encontradas numa base de dados. Desde o seu surgimento em 1994, que diversas variantes do *Apriori* têm sido propostas, por forma a melhorar a eficiência computacional do mesmo. Desta forma, surgiram os algoritmos *AprioryTid* e *AprioriHybrid* (combinação do *Apriori* com o *AprioryTid*) que permitiram melhorar o desempenho do *Apriori*. Em 1995, o algoritmo OPUS (WEBB, 1995) veio assumir-se como uma alternativa à utilização do algoritmo *Apriori* (e suas variantes). Este algoritmo apresenta melhorias em termos de tempo de execução, quando aplicado a determinados problemas, normalmente em domínios não relacionados com o *market basket analysis* (WEBB, 2000).

1.1 Geração de regras

O algoritmo *Apriori* divide a descoberta de regras de associação em dois subproblemas (AGRAWAL et al., 1993):

- descoberta de todos os conjuntos de itens frequentes;
- geração das regras de associação (utilizando os conjuntos de itens frequentes descobertos).

De entre estes passos, o primeiro é mais exigente sob o ponto de vista computacional, e tem merecido especial atenção por parte da comunidade de *data mining* (ZHENJIANG et al, 2001), levando ao surgimento de diferentes algoritmos cujo principal objectivo reside na melhoria da eficiência computacional do *Apriori*.

No esquema seguinte (figura 7) está representado o processo de descoberta das regras de associação, a partir de uma base de dados de transacções e que obedecem a um suporte mínimo de 50%.

Base de dados de transacções

Transacção	Itens (produtos)
1	Arroz, Azeite, Massas
2	Arroz, Massas
3	Arroz, Batatas
4	Azeite, Feijão

Conjuntos frequentes

Conjunto frequente	Suporte
Arroz	3/4 = 75%
Azeite	2/4 = 50%
Massas	2/4 = 50%
Arroz, Massas	2/4 = 50%

Regras

Regra	Suporte	Confiança
Arroz → Massas	2/4 = 50%	(2/4) / (3/4) = 67%
Massas → Arroz	2/4 = 50%	(2/4) / (2/4) = 100%

Figura 2 Fases de processamento do algoritmo *Apriori* para a produção de regras de associação.

Observando o esquema anterior, verifica-se que processo de obtenção de regras de associação inicia-se pela identificação dos itens ou produtos que fazem parte de cada uma das transacções. Após esta fase, o algoritmo de geração de regras determina e selecciona as associações de itens que ocorrem com mais frequência, nas transacções.

Com base nos conjuntos de itens frequentes, geram-se todas as regras de associação que obedeçam aos valores mínimos de suporte e de confiança. Tratando-se de bases de dados com centenas de milhares ou milhões de transacções, os algoritmos irão produzir também milhares de regras de associação, dependendo também do número de itens diferentes existentes.

A geração das regras de associação pode, para grande volume de dados, pode originar um número de regras geradas tão elevado que facilmente ultrapassa o limiar da

legibilidade humana. Na prática, o número de regras produzidas pelos algoritmos pode ser tão elevado que se fica perante um novo problema de *data mining*: extrair conhecimento das regras de associação produzidas (TOIVONEN et al., 1995). Também nesta área, diversos investigadores desenvolveram métodos que procuram resolver este problema. O objectivo desses métodos é o de encontrar as "melhores regras", as "regras óptimas" ou as "regras mais interessantes", obtidas a partir de uma base de dados, de acordo com uma variedade de medidas tais como confiança, suporte, *gain*, *laplace*, *lift, conviction* e muitas outras.

1.2 Suporte e confiança

Cada uma das regras de associação, gerada pelos algoritmos de descoberta de regras, representa um padrão de uma base de dados. Na geração das regras de associação, os algoritmos procuram regras que satisfaçam determinadas condições. Essas condições correspondem, normalmente, a um valor mínimo de suporte e confiança. O suporte de uma determinada regra mede a frequência com que esse padrão aparece na base. A confiança corresponde a um valor de correlação entre os itens que formam esse padrão. Para evitar que se gere um número de regras quase tão elevado quanto o número de transacções em análise, geralmente, condicionam-se as regras para um suporte baixo e uma confiança elevada.

Considere-se um outro exemplo prático de regra de associação a afirmação de que "40% das compras de cerveja também incluem *Seven-up* e 10% de todas as vendas incluem ambos os produtos (itens)". O valor 40% designa-se por confiança da regra e o valor 10% corresponde ao suporte da regra.

Pode definir-se o suporte *s* de uma regra do seguinte modo:

> Considerando o conjunto de transacções D de uma base de dados, o suporte s de uma regra $A \rightarrow B$ corresponde ao número de transacções T, existente em D, que contêm todos os itens de $A \cup B$.

De forma semelhante define-se formalmente a confiança *c* de uma regra:

> Por confiança *c* de uma regra $A \rightarrow B$, entende-se a proporção entre o número de transacções que contêm os itens de $A \cup B$ e o número de transacções que contêm os itens de *A*.

De acordo com as definições anteriores, verifica-se ainda que, o suporte de uma regra $A \rightarrow B$ pode ser entendido como a probabilidade de que todos os itens de $A \cup B$ estejam presentes numa transacção. De igual modo, a confiança corresponde à probabilidade condicional observada de encontrar o conjunto de itens *B*, tendo encontrado o conjunto de itens *A*.

Uma das preocupações na produção de regras de associação consiste na selecção de regras que satisfaçam determinados níveis mínimos de suporte e confiança exigidos pelo utilizador[1]. Normalmente, exige-se um suporte baixo e uma confiança alta para valores mínimos. O suporte mínimo é exigido por forma a não obter demasiadas regras (dependendo da dimensão da base de dados pode tratar-se de um universo de dezenas ou centenas de milhares de regras). Definindo-se um suporte mínimo restringe-se, assim, a quantidade de regras no output do algoritmo de regras de associação. Interessa garantir uma confiança alta, em função do valor esperado à priori, para que exista uma elevada coesão entre os itens analisados. Uma confiança baixa não reflectiria qualquer padrão de comportamento.

Nas figuras seguintes (figura 3 e 4), apresenta-se, para o caso particular do conjunto de dados *German Credit Data*[2], a relação entre os valores mínimos de suporte e confiança, considerados pelos algoritmos, para a geração de regras de associação. Para ambos os casos estabeleceu-se um limite máximo de 1 000 regras produzidas.

[1] Utilizador: pessoa que tem a seu cargo o trabalho manual inerente a qualquer elemento de um sistema informático.

[2] *German Credit Data*: conjunto de dados relativos a aprovação de empréstimos bancários, frequentemente utilizados na avaliação e aferição de aplicações de *data mining*.

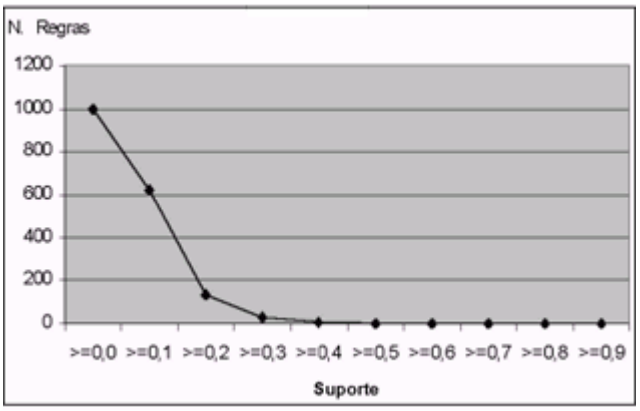


Figura 3 Relação entre o número de regras geradas e diferentes valores de suporte mínimo.

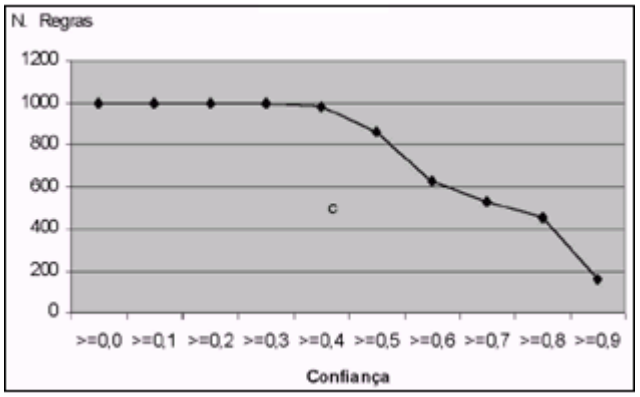


Figura 4 Relação entre o número de regras geradas e diferentes valores de confiança mínima.

A observação do gráfico da figura 3 mostra como pode evoluir o número de regras de associação produzidas, à medida que se exige um suporte mínimo maior (estabilizando a partir de 0,4). O aumento no valor mínimo de confiança, representada no gráfico da figura 4, tem também repercussões no número de regras geradas, embora este efeito se comece a fazer sentir para valores de confiança superiores a 0,4. Para valores de suporte muito baixos, pequenas variações no suporte mínimo podem originar grandes variações no número de regras produzidas.

O suporte e a confiança são as medidas de interesse universalmente mais utilizadas na geração de regras de associação. Fazendo-se variar o valor destas duas medidas, o

número de regras produzidas pode também variar de forma expressiva. Aumentando-se o suporte mínimo a partir do qual uma regra é aceite, diminuirá o número de regras geradas pelo algoritmo, uma vez que se exige que um determinado conjunto de itens (constituinte de uma regra) exista num maior número de registos da base de dados. Do mesmo modo, fazendo-se aumentar a confiança mínima das regras aceites, o algoritmo irá gerar menos regras porque se exige uma proporção maior entre o número de transacções que contém o conjunto dos itens que constituem uma regra e o número de transacções que contêm os itens do antecedente.

1.3 Outras medidas de interesse

Apesar da reconhecida importância do suporte e confiança, como forma de caracterizar e avaliar o interesse das regras de associação, importa referir outras medidas que podem ajudar nessa caracterização. Estas medidas podem ainda ser utilizadas como meio restringir a produção de determinadas regras, tal como acontece com o suporte e a confiança. Desta forma podem ser definidos *rankings*, que permitam ordenar as regras segundo uma determinada medida. De seguida serão enumeradas e descritas algumas dessas medidas: *interest (ou lift)*, *coverage*, *leverage* e *conviction*.

Interest ou Lift

O *Interest* ou *Lift* de uma regra $A \rightarrow B$ corresponde ao quociente entre a probabilidade conjunta (de A e B) observada e a probabilidade conjunta sob independência (BRIN et al., 1997b). Esta medida pode ser representada da seguinte forma:

$interest(A \rightarrow B) = P(A \cap B) / (P(A) * P(B))$, que é o mesmo que ter:

$interest(A \rightarrow B) = suporte(A , B) / suporte(A) * suporte(B)$

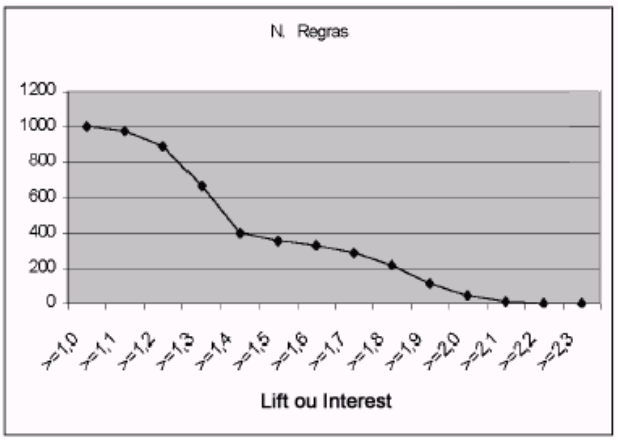


Figura 5 Relação entre o número de regras geradas e diferentes valores de *Lift* ou *Interest*.

Como se pode concluir pela observação do gráfico da figura 5 também no caso desta medida o número de regras produzidas varia significativamente se houver um acréscimo no valor de *Interest* considerado.

Coverage

A medida *coverage* define a proporção de exemplos cobertos pelos itens que compõem o antecedente da regra (*Rulequest Research*). Assim sendo, pode definir-se esta medida da seguinte forma:

$$coverage(A \rightarrow B) = \text{suporte}(A)$$

Leverage

A medida *leverege* define a diferença entre a proporção de exemplos cobertos, simultaneamente, pelo antecedente e pelo consequente da regra e a proporção de exemplos que seriam cobertos se o antecedente e o consequente fossem independentes (*Rulequest Research*).

$$leverage\ (A \rightarrow B) = \text{suporte}\ (A \cap B) - (\ \text{suporte}(A) * \text{suporte}(B)\ )$$

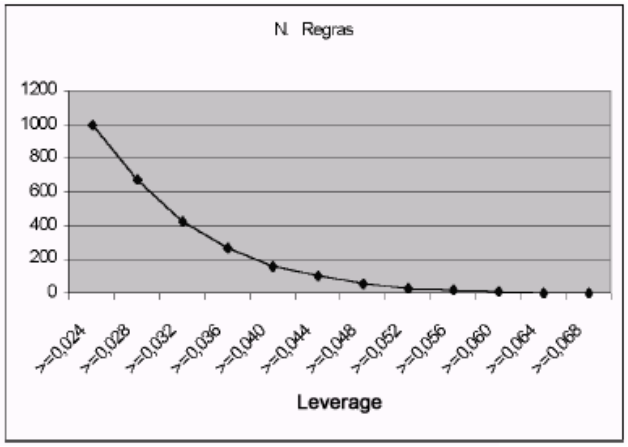


Figura 6 Relação entre o número de regras geradas e diferentes valores de *Leverage*.

Pode observar-se, uma vez mais (figura 6), que a definição de um valor mínimo de *leverage* tem implicações no número de regras de associação geradas pelos algoritmos.

Conviction

A medida *conviction* permite medir a independência do antecedente *A*, face ao consequente *B* (BRIN et al., 1997b). Trata-se de uma medida unidireccional, isto é, a *conviction*($A \rightarrow B$) é diferente da *conviction*($B \rightarrow A$). Assim, esta medida pode representar-se do seguinte modo:

$$conviction(A \rightarrow B) = P(A \cap \neg B) / (P(A) * P(\neg B))$$

O valor de *conviction* pode variar entre 0 e $+\infty$, apresentando o valor 1 quando os conjuntos *A* e *B* são independentes.

Teste do χ^2 (qui-quadrado)

A utilização do teste estatístico do qui-quadrado pode também ser importante como forma de medir a correlação entre o antecedente e os consequente de uma regra (LIU et al., 1999b). Este teste permite também indicar a direcção da correlação (positiva, negativa ou independência). Este teste baseia-se na comparação das frequências observadas com as correspondentes frequências esperadas. Quanto mais próximas estiverem estas duas frequências, maior será a probabilidade de se tratarem de casos

independentes. Desta forma, o teste do χ^2 é utilizado para testar a significância do desvio face aos valores esperados. Considerando f_0 uma frequência observada e f uma frequência esperada, o valor do χ^2 é definido do seguinte modo:

$$\chi^2 = \sum \frac{(f - f_0)^2}{f}$$

Um valor de $\chi^2 = 0$ significa que os atributos (antecedente e consequente de uma regra) são estatisticamente independentes. No caso de não se poder confirmar a independência dos atributos, pode medir-se a correlação entre eles, considerando o valore esperado e observado do suporte da regra. Considera-se, assim, que:

- se Sup0($A \rightarrow B$) > Sup($A \rightarrow B$) então existe uma correlação positiva entre *A* e *B*
- se Sup0($A \rightarrow B$) < Sup($A \rightarrow B$) então existe uma correlação negativa entre *A* e *B* (considera-se uma regra desinteressante)

A maior parte das medidas acima referidas tem por finalidade aferir da independência entre o antecedente e o consequente da regra. Desta forma, podem ser utilizadas como forma de filtrar as regras de associação que obedeçam a determinados valores de independência, permitindo reduzir o número de regras produzidas. Pode-se, por exemplo, definir que se pretende obter regras que possuam um valor de *leverage* superior a 0,03. De forma semelhante, pode restringir-se o conjunto de regras de associação produzidas às regras que possuam um valor mínimo para o *lift* de 1,4.

2 Pós-processamento de regras de associação

Pode-se pensar que o processo de geração de regras de associação assim como a sua compreensão são uma tarefa simples. Se este raciocínio é verdadeiro para pequenos conjuntos de regras, tal já não acontece quando se procura analisar um grande conjunto de regras de associação. Em seguida, serão abordados alguns estudos levados a cabo por diversos investigadores, cujo objectivo principal era o de facilitar a interpretação de um grande número de regras de associação. Foram propostos diferentes métodos: eliminação de regras redundantes, identificação das regras mais interessantes ou menos interessantes, utilização de técnicas de visualização e recurso a sistemas de gestão de bases de dados.

2.1 Resumo e agrupamento

Uma das formas encontradas para reduzir o número de regras produzidas pelos algoritmos de geração de regras de associação foi o recurso à "poda" (redução) de regras redundantes (TOIVONEN et al., 1995 e LIU et al., 1999b). Embora recorrendo à "poda" de regras ou *pruning* por forma a reduzir o número de regras descobertas, cada um dos investigadores, acima referidos, utilizaram métodos diferentes de realizar essa redução do número de regras, obtendo também resultados diferenciados. Toivonen recorreu à noção de cobertura de uma regra (*rule cover*) por forma a eliminar regras que não oferecessem informação adicional. A cobertura de uma regra envolve a identificação de regras que descrevam os mesmos registos de uma base de dados, ou seja, regras cujos itens que as constituem sejam mais específicos (mais desagregados) mas não possuam um valor de confiança mais elevado. As *rule covers* significam, no essencial, a aplicação de filtros a um conjunto de regras. Neste estudo, Toivonen apresenta um exemplo que pode ajudar a compreender esta ideia. Considerando uma base de dados de alunos que escolheram determinados cursos tecnológicos, duas das regras geradas pelo *Apriori* foram:

1) Curso de C, Curso de Bases de Dados → Curso de Comunicação de Dados (c=0,90; s=0,02)

2) Curso de C, Curso de Bases de Dados, Curso de Utilização do Computador → Curso de Comunicação de Dados (c=0,90; s=0,01)

Segundo este investigador, a segunda regra é redundante, sendo mais específica que a primeira e não produzindo informação adicional, uma vez que a confiança mantém-se inalterada e o suporte da segunda reduz-se para metade. Considera-se que a primeira regra "cobre" a segunda regra. Constatando que as regras obtidas por este método continuavam a ser numerosas, Toivonen complementou o estudo propondo uma ordenação por ordem de interesse das regras e posterior agrupamento das mesmas, recorrendo a um método estatístico de *Clusters*. Este investigador, apesar de não aprofundar este método, defende que uma possível forma de definir estes *Clusters* será recorrendo à distância entre as diversas regras de associação. Toivonen propõe que a distância *d* entre duas regras de associação seja medida da seguinte forma:

$$d(A \rightarrow B, C \rightarrow D) = |(m(AB) \cup m(CD)) \setminus m(ACD)|$$
$$= |(m(AB) + m(CD)| - 2|m(ACD)|$$

em que $m(AB)$ identifica o número de registos da base de dados que possuem os itens constituintes dos conjuntos *A* e *B*.

Desta forma, Toivonen consegue facilitar o processo de análise de regras de associação, não só através do recurso às *rule covers* como também da posterior ordenação e *clustering* das regras.

Liu, utilizando também por base a redução de regras redundantes, recorre ao método do Qui-quadrado para identificar as regras com associações mais significativas, procurando obter posteriormente um conjunto especial de regras que resume as restantes. A este conjunto especial de regras, Liu designa por regras DS (*Direction Setting rules*), ou seja, regras que procuram orientar a análise do utilizador para determinadas direcções

do espaço de regras. O esquema seguinte (figura 7) representa a técnica proposta, constituída por dois passos principais, "poda" (*pruning*) e resumo (*summarization*).

Fonte: LIU et al., 1999b.

Figura 7 Fases que constituem a técnica introduzida por Liu.

Nesse estudo, é utilizada a correlação estatística (recorrendo ao teste do Qui-quadrado), em vez da confiança mínima, como base para encontrar as regras que melhor representem as relações entre conjuntos de itens (regras mais significativas). À semelhança do Toivonen, também Liu prefere as regras simples e gerais às regras mais específicas. As regras DS, obtidas após uma redução de regras inicial, correspondem a um subconjunto de regras que pretende sumariar o conjunto de regras "podadas". Este subconjunto de regras identifica as regras de associação mais significativas, que servem de ponto de partida para o restante espaço de regras (*non-DS rules*).

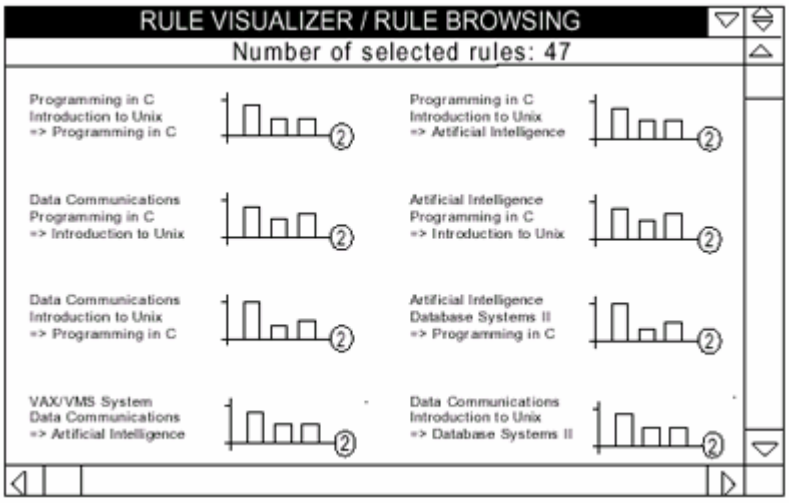
2.2 Visualização

Uma área que também tem tido algum desenvolvimento, no que respeita à sua utilização nos processos de *data mining* e, mais concretamente, na interpretação de regras de associação, é a visual ou gráfica. Klemettinen, em conjunto com outros investigadores, elaborou um sistema de visualização de regras de associação, o *Rule Visualizer* (KLEMETTINEN, M. et al*,* 1994). Este sistema aplicava o conceito de *templates* ou de filtros que será analisado no ponto seguinte. O *Rule Visualizer*, desenvolvido com o intuito de apoiar a análise do utilizador da descoberta de regras de associação, era composto por três componentes distintas: *Rule Selection*, *Rule Browsing* e *Rule Graph*. A primeira componente diz respeito, como irá ser referido no ponto seguinte, à utilização de filtros por forma a efectuar uma selecção das regras interessantes. As

outras duas componentes permitiam representar, graficamente, as regras de associação interessantes.

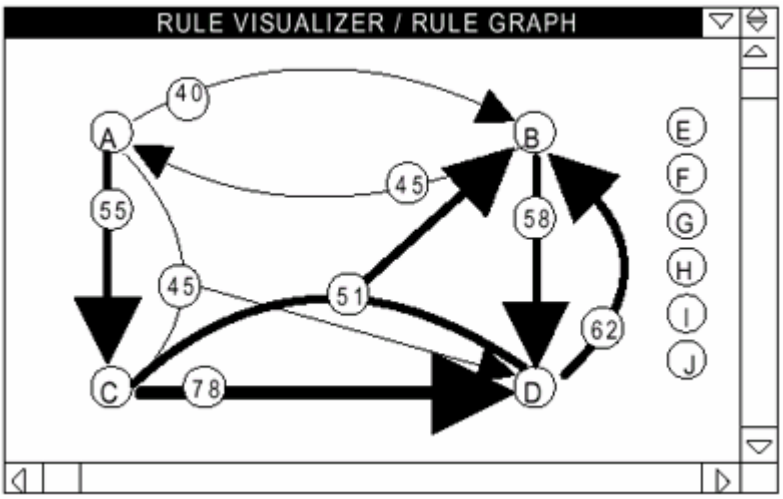
A componente *Rule Browsing* (figura 8) permite ao utilizador observar o espaço de regras, através da visualização, sob a forma de gráficos de barras, das regras interessantes. Nos gráficos, são definidas as medidas de suporte, confiança e a medida *commonness* que representa uma forma de relacionar as duas medidas anteriores, ou seja, definindo uma das medidas em função da outra.

Por seu lado, o *Rule Graph* (figura 9) permite uma visualização das regras de associação sob a forma de grafo, em que os nós representam itens e a os arcos que ligam os atributos representam a confiança e o suporte. A largura dos arcos permite distinguir diferentes valores para a confiança e suporte.



Fonte: KLEMETTINEN, M. et al, 1994.

Figura 8 Regras de associação visualizadas na componente Rule Browsing do Rule Visualizer.



Fonte: KLEMETTINEN, M. et al, 1994.

Figura 9 Regras de associação visualizadas na componente Rule Graph do Rule visualizer.

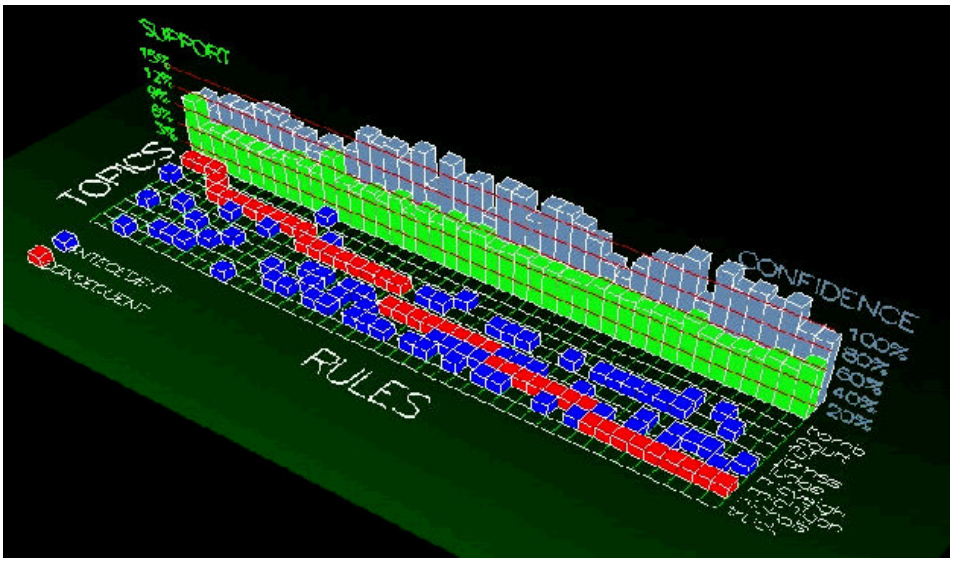
Em 1999, surgiu um trabalho que propõe um novo método para ajudar o utilizador a explorar as regras descobertas (LIU, B. et al., 1999a). Este novo método dividiu o problema da exploração das regras de associação em duas componentes: a componente de análise das regras interessantes e a componente de visualização. Liu aponta como principal dificuldade, na selecção das regras interessantes, o facto de se tratar de uma tarefa revestida de subjectividade, uma vez que o interesse por determinadas regras depende do conhecimento do utilizador no domínio em análise.

Após esta análise prévia definida pelo utilizador, é proposta uma forma visual de apresentar as regras potencialmente interessantes. O conceito inerente a esta componente visual é o de aproveitar as capacidades visuais humanas para identificar, mais rapidamente, as regras realmente interessantes. Esta componente consiste na realização de um interface composto por quatro módulos ou funções distintas. O primeiro módulo permite criar classes ou hierarquias de itens (*class hierarchy builder*), o que, segundo Liu, pode ser importante para a identificação de itens interessantes. Os restantes módulos (*GI viewer, RPC viewer e PK viewer*) possibilitam a visualização de associações entre regras e entre classes de itens (envolvidos nas regras visualizadas).

Em 2000, com a colaboração de Liu, a técnica das regras DS foi implementada num interface[3] de páginas *web*[4] que permitiam ajudar o utilizador a percorrer o espaço de regras, partindo das regras DS e percorrendo as regras *non-DS rules* (MA et al., 2000). O interface proposto procurava aliar as potencialidades da "poda" e resumo das regras de associação às potencialidades da Internet como meio privilegiado de comunicação e como ambiente familiar para percorrer o espaço de regras. O sistema DS-WEB utilizava o mesmo tipo de abordagem que o sistema apresentado neste trabalho. Em comum, o DS-WEB e o PEAR têm o objectivo de pós-processamento de um grande número de regras de associação recorrendo à navegação através de páginas *web* e à visualização gráfica. O sistema DS-WEB baseia-se na utilização de um conjunto mais reduzido de regras (*DS rules*), partindo depois para análise a variantes deste tipo de regras. O PEAR, no entanto, possibilita a aplicação de operadores a qualquer tipo de regra, incluindo regras do tipo *DS-rules*. O conjunto de operadores apresentados neste trabalho é baseado em simples propriedades matemáticas de conjuntos e possuem uma semântica bastante intuitiva e clara.

Um outro estudo na área da visualização de regras de associação teve como principal objectivo o desenvolvimento de uma técnica de representação gráfica de regras de associação (WONG et al., 1999). Com interesse para a visualização de regras, foram identificados cinco parâmetros fundamentais: itens constituintes dos antecedentes, itens constituintes dos consequentes, associações entre antecedentes e consequentes, suporte das regras e confiança das regras. Neste estudo, várias técnicas anteriores de representação gráficas foram analisadas e comparadas. O principal problema encontrado nas técnicas analisadas tinha que ver com um certo limite no número de regras representadas, pelo que a preocupação de Wong focalizou-se na resolução desta questão.

[3] Interface: dispositivo gráfico de ligação entre os sistemas informáticos e o utilizador que permite apresentar os dados e as funções de um programa.
[4] *Web*: palavra inglesa que identifica a forma multimédia de aceder à Internet: a *wold wide web*. designam-se frequentemente por páginas *web* os documentos escritos em linguagem HTML e que estão disponíveis na Internet.



Fonte: WONG, P. C. et al., 1999.

Figura 10 Visualização de regras de associação ordenadas pelo valor de confiança.

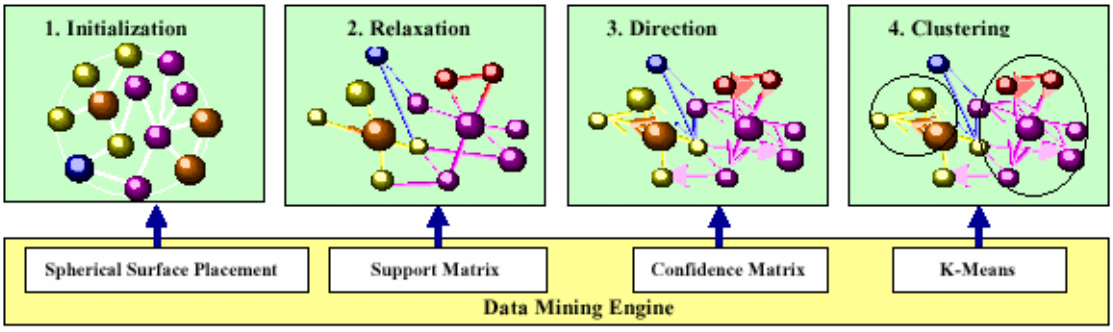
A proposta apresentada evidenciava uma abordagem de visualização do tipo regra-item (figura 10), cujas principais vantagens apontadas, em relação a outras representações gráficas, eram:

- virtualmente não existir limite superior no número de itens do antecedente;
- ser possível analisar, em simultâneo, a distribuição das regras (eixo horizontal) bem como os itens respectivos (eixo vertical);
- não existir problemas com sobreposições de barras no gráfico;
- ser suficiente, para a interpretação do gráfico, a possibilidade de aproximar ou distanciar determinados pormenores.

Segundo a experiência realizada pelos investigadores, esta técnica de visualização funcionou bem, mesmo com algumas centenas de regras de associação.

Um estudo mais recente procurou aplicar técnicas de visualização de regras de associação ao comércio electrónico (HAO, M. C. et al., 2000). Os investigadores propuseram a criação de um sistema de visualização de associações e relações entre os itens (produtos), a partir de um grande volume de dados provenientes de transacções por

comércio electrónico. Este sistema foi baptizado de DAV (*Directed Association Visualization*). As limitações com as representações gráficas habituais (em matrizes que exprimem as relações entre itens) levaram a que no sistema DAV fosse implementada uma técnica de representação gráfica de associações baseada em grafos. Este sistema, implementado na linguagem de programação JAVA, é constituído por quatro componentes básicas – *initialization*, *relaxation*, *direction* e *clustering* (figura 11). Utilizando estes componentes, o agente decisor pode fazer uma sua análise apoiada em gráficos que evidenciam os grupos de produtos que exibem fortes relações entre si.



Fonte: HAO, M. C. et al., 2000.

Figura 11 Arquitectura dos componentes que constituem o sistema DAV.

VizWiz é a designação não-oficial para um visualizador interactivo de modelos de *data mining* em formato PMML (WETTSHERECK, 2002). Este visualizador foi implementado em linguagem Java e permite visualizar graficamente, tanto regras de associação como muitos outros modelos de *data mining*. A filosofia do VizWiz para visualizar regras de associação assenta na representação de uma lista de regras de associação que cumpram determinados valores de suporte e confiança definidos pelo utilizador. Os valores de suporte e confiança de cada regra são representados sob a forma de barras coloridas.

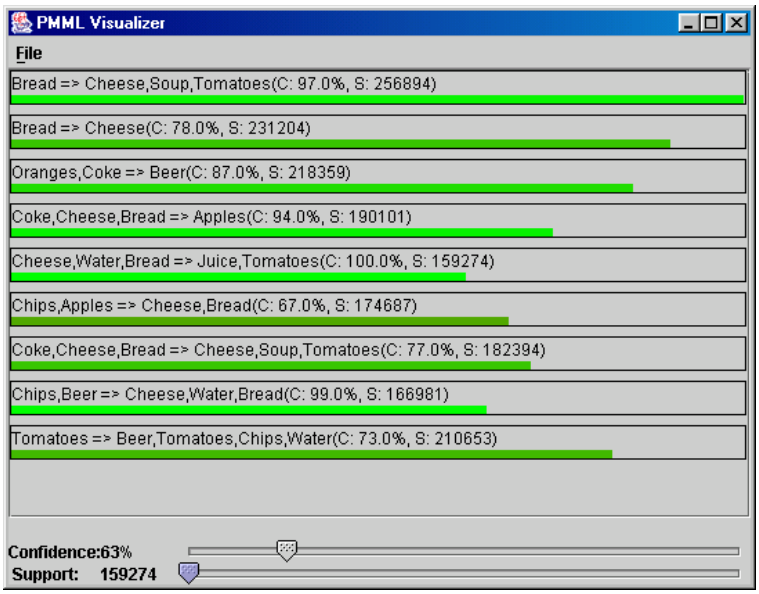


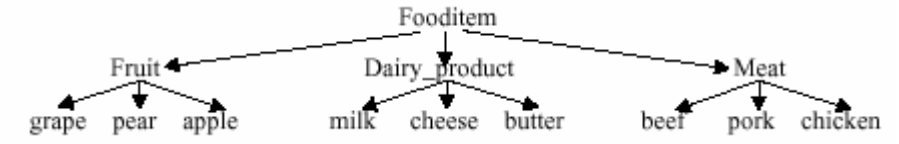
Figura 12 Aspecto do interface do VizWiz.

A figura anterior apresenta um aspecto do visualizador VizWiz com uma lista de regras de associação. Este visualizador pode ser utilizado directamente no sistema operativo ou integrado numa página no *web browser*.

2.3 Medidas de interesse e regras interessantes

Muitos autores têm procurado encontrar propriedades, nas regras de associação, que permitam medir a sua importância (medidas de interesse). Um dos estudos abordados no ponto anterior deste texto (LIU et al., 1999a) propunha a utilização de um interface de visualização no apoio à descoberta de regras interessantes. No entanto, para que esta visualização fosse mais eficaz, era necessário uma outra componente que fizesse a análise das regras interessantes. O interesse de uma regra é, segundo o autor, algo bastante subjectivo, que depende dos objectivos de quem a estuda. Desta forma, uma regra interessante para um utilizador pode não ser interessante para outro. A proposta para a componente de análise de regras interessantes baseava-se, assim, na utilização do conhecimento prévio do utilizador, para identificação de vários tipos de regras potencialmente interessantes. Nesta componente, foi definida uma linguagem de especificação de conhecimento, segundo a qual o utilizador classifica e define um conjunto de associações de itens seguindo a sua experiência e saber. Os itens que

compõem as regras são também agrupados em classes hierárquicas. O esquema representado na figura seguinte exemplifica este tipo de agrupamento.



Fonte: LIU et al., 1999a.

Figura 13 Exemplo de uma hierarquias de produtos.

Para a especificação de conhecimento foram definidos três tipos de associações:

- *General Impression* (GI): representa um sentimento vago, de que existe um conjunto de associações entre algumas classes de itens, sendo que o utilizador possui poucas certezas;
- *Reasonably Precise Concept* (RPC): o utilizador tem algumas certezas em determinadas associações de classes de itens e consegue identificar as direcções dessas associações;
- *Precise Knowledge* (PK): quando existe uma grande certeza acerca de determinadas associações de classes de itens.

Perante estas especificações ou "crenças" do utilizador do algoritmo de descoberta de regras, estas seriam então classificadas. Neste sentido, foi criada uma nomenclatura que permitiria classificar as regras em quatro tipos diferentes:

- *Conforming rules*: regras consistentes ou em conformidade com as especificações iniciais;
- *Unexpected consequent rules*: regras cujos consequentes contrariam as "crenças" do utilizador. São consideradas regras potencialmente interessantes;
- *Unexpected condition rules*: regras cujos antecedentes contrariam as "crenças" do utilizador. Podem também conduzir a regras desconhecidas e, por isso mesmo, interessantes;

- *Both-side unexpected rules*: regras cujos antecedentes ou consequentes não foram mencionados nas especificações iniciais, pelo que permite descobrir novos espaços de regras;

Esta nomenclatura permitia produzir, assim, uma lista de regras de associação, ordenadas por ordem de interesse, dentro de cada uma das quatro classificações.

Em 1999, Sahar, propôs uma abordagem quase inversa para a resolução do problema de selecção das regras de associação mais interessantes: a eliminação das regras desinteressantes (SAHAR, S., 1999). Ao invés de tentar identificar as regras potencialmente interessantes, este investigador procura definir as regras desinteressantes, eliminando-as para que o conjunto de regras final seja constituído pelas regras interessantes. O autor começou por separar, à semelhança de outros investigadores (SILBERSCHATZ e TUZHILIN, 1996), as medidas de interesse objectivas das medidas de interesse subjectivas. As medidas de interesse objectivas são aquelas medidas que dependem exclusivamente da estrutura dos dados e dos padrões retirados dos mesmos, enquanto as medidas de interesse subjectivas também dependem de necessidades específicas e do conhecimento prévio do utilizador. Estes investigadores concentraram-se nas medidas de interesse subjectivas, porque são, as que em último caso, interessam mais a um utilizador específico. Para determinar o que é subjectivamente interessante, torna-se necessário incorporar, de alguma forma, o conhecimento do utilizador no sistema proposto. A forma encontrada para transmitir ao sistema este conhecimento consistiu numa classificação iterativa de regras desinteressantes efectuada pelo utilizador. Para cada regra apresentada para classificação, é solicitado ao utilizador que indique se a regra é verdadeira e se há interesse (definir o tipo de interesse) em alguma regra que possa ser semelhante à regra em causa (da mesma família). Indicar que uma regra é ou não verdadeira implica que se pode excluir regras "desinteressantes" ou de "senso comum". Para caracterizar o interesse das regras, foram propostas quatro categorias de classificação de regras, que permitem definir se a regra é ou não verdadeira e se possui ou não interesse. Essas quatro categorias correspondem às seguintes designações:

- *True-Not-Interesting* (TNI) – classificam regras que possuem informação redundante, e, por esse motivo, são consideradas desinteressantes;
- *Not-True-Interesting* (NTI) – designam regras que, não correspondendo a uma certeza absoluta, podem ser consideradas interessantes para a análise em questão;
- *Not-true-Not-Interesting* (NTNI) – classificam regras que, não correspondendo a uma certeza absoluta, podem ser consideradas desinteressantes para o problema em análise;
- *True-Interesting* (TI) - correspondem a regras que, sendo certezas absolutas, são classificadas pelo analista como interessantes.

O estudo das regras que podem suscitar maior interesse, para os analistas das regras de associação, continua a ser uma área em franco desenvolvimento. A prová-lo, surge mais um investigador, Hussain, que propõe definir medidas relativas de estimar o interesse de uma regra (HUSSAIN et al., 1999). Para este investigador, uma regra interessante está sempre relacionada com o conhecimento e necessidades de quem a está a analisar, pelo que há sempre uma subjectividade inerente a este tipo de classificação. Contudo, pelo ponto de vista objectivo, também se pode identificar uma regra interessante, recorrendo a medidas concretas, tais como a confiança e o suporte. Para a classificação de regras interessantes, Hussain considerou os seguintes factores:

- Suporte (*support*): mede a frequência com que os itens constituintes da regra aparecem no conjunto de transacções;
- Confiança (*confidence*): corresponde a um valor de correlação entre os itens que formam a regra;
- Senso comum (*common sense*): uma regra de acordo com o "senso comum" corresponde a uma regra com valores de suporte e confiança altos;
- Fiabilidade (*reliability*): explica o significado estatístico da regra. As regras com maior significado estatístico alta são mais fiáveis;
- Aplicabilidade (*actionability*): mede a capacidade de uma regra originar uma acção/decisão. As regras mais aplicáveis correspondem a regras fiáveis que podem ser aplicadas a determinado domínio em análise;

- Novidade (*novelty*): regras que são completamente novas para o utilizador. Não são regras necessariamente aplicáveis ou fiáveis;
- Surpresa (*unexpectedness*): representam uma regra surpresa porque vão contra o que seria de esperar. Uma regra pode ser inesperada face ao que o utilizador pensa ou inesperada face às regras de senso comum.

Dos factores apresentados, este estudo destaca os três últimos (aplicabilidade, novidade e surpresa), uma vez que são factores subjectivos de análise de uma regra. No entanto, Hussain defende que, apesar da sua aparente subjectividade, estes factores estão intimamente ligados com os factores objectivos assinalados acima: suporte, confiança, senso comum e fiabilidade. Este autor considera que os factores aplicabilidade, novidade e surpresa não são mutuamente exclusivos, pelo que as regras interessantes podem ser classificadas de acordo com a tabela seguinte.

Tabela 1 Categorias de regras interessantes.

	Inesperada	**Esperada**	**Nova**
Aplicável	**++ interessante**	- interessante	+ interessante
Inaplicável	- interessante	desinteressante	desinteressante

Fonte: adaptada de HUSSAIN et al., 1999.

Perante a identificação destas categorias, foi possível definir que, para ser possível encontrar uma medida objectiva que "medisse" o interesse de uma regra, teria que se estimar a aplicabilidade e surpresa da mesma. Uma análise mais pormenorizada destes resultados levou o autor a defender que, para se classificar uma regra de aplicável, teria que existir a intervenção do utilizador. A este tipo de classificação está inerente, por isso, uma subjectividade de quem analisa as regras de associação. No entanto, para que uma regra seja classificada de inesperada, seria suficiente considerá-la contrária às regras de senso comum (regras com altos valores de suporte e confiança). Desta forma, este investigador propôs que a surpresa de uma regra pudesse ser considerada uma medida de interesse, para a classificação de regras de associação. A surpresa de uma regra pode ser observada sob duas formas distintas: suporte inesperado e confiança inesperada. Uma regra que possua, simultaneamente, um suporte inesperado e uma

confiança inesperada deverá ser mais interessante do que uma regra que apenas possua uma das medidas inesperadas.

Klemettinen, em conjunto com outros investigadores, propôs a utilização de *template rules* (filtros de regras), por forma a descrever as regras que interessa obter (KLEMETTINEN, M. et al*,* 1994). O utilizador deverá fornecer informação adicional acerca da estrutura dos dados. Os *templates* permitem distinguir, previamente, o tipo de regras que se considera interessantes das que se considera desinteressantes. Para esta distinção, o autor definiu dois *templates*: *inclusive template* e *restrictive template*. Para que uma determinada regra se considere interessante, deve "encaixar" num dos *inclusive templates*. Se, por outro lado, uma regra "encaixa" num dos *restrictive templates*, a regra será considerada desinteressante. Para que estes *templates* possam ser aplicados, torna-se necessário proceder a uma classificação prévia, por forma a definir uma hierarquia de itens (classes de itens). Os *templates* permitem filtrar, segundo as classes definidas também pelo utilizador, as regras produzidas pelos algoritmos.

2.4 Recurso a bases de dados

A popularidade que os sistemas de gestão de bases de dados têm vindo a assumir, motivada, em grande parte, pela utilização da linguagem *Structured Query Language*[5] (SQL), levou a que diversos investigadores procurassem tirar partido das capacidades destes sistemas na área do *data mining*.

Em 1996, um estudo sobre um operador, baseado em SQL, para a descoberta de regras de associação, foi levado a cabo por Rosa Meo (MEO, R. et al, 1996). Este operador, designado por MINE RULE, permite a descoberta de regras de associação, a partir de uma tabela de transacções. O MINE RULE representa uma forma de, recorrendo apenas a uma linguagem baseada em SQL e exprimindo condições definidas pelo utilizador, extrair um conjunto de regras de associação. Nas figuras seguintes, pode observar-se um

[5] *Structured Query Language* (SQL): linguagem standard de manipulação e definição de bases de dados, reconhecida pelos principais sistemas de gestão de bases de dados.

conjunto de instruções com determinadas condições (instrução MINE RULE e respectivas cláusulas) que, partindo de uma tabela de transacções, permite extrair as regras de associação.

Tabela 2 Exemplo de uma base de dados de transacções, agrupada por cliente.

tr.	*customer*	*item*	*date*	*price*	*q.ty*
1	$customer_1$	ski_pants	12/17/95	140	1
	$customer_1$	hiking_boots	12/17/95	180	1
2	$customer_2$	col_shirts	12/18/95	25	2
	$customer_2$	brown_boots	12/18/95	150	1
	$customer_2$	jackets	12/18/95	300	1
3	$customer_1$	jackets	12/18/95	300	1
4	$customer_2$	col_shirts	12/19/95	25	3
	$customer_2$	jackets	12/19/95	300	2

Fonte: MEO, R. et al, 1996.

```
MINE RULE SimpleAssociations AS
SELECT DISTINCT 1..n item AS BODY,
                1..1 item AS HEAD,
                SUPPORT, CONFIDENCE
FROM Purchase
GROUP BY transaction
EXTRACTING RULES WITH SUPPORT: 0.1,
                      CONFIDENCE: 0.2
```

Fonte: MEO, R. et al, 1996.

Figura 14 Instrução MINE RULE, e respectivas cláusulas, para extrair regras de associação.

Tabela 3 Tabela de dados *SimpleAssociations* que contém as regras de associação que verificam as condições explícitas nas instruções SQL.

BODY	*HEAD*	*S.*	*C.*
{ski_pants}	{hiking_boots}	0.25	1
{hiking_boots}	{ski_pants}	0.25	1
{col_shirts}	{brown_boots}	0.25	0.5
{col_shirts}	{jackets}	0.5	1
{brown_boots}	{col_shirts}	0.25	0.5
{brown_boots}	{jackets}	0.25	1
{jackets}	{col_shirts}	0.5	0.66
{jackets}	{brown_boots}	0.25	0.33
{col_shirts,brown_boots}	{jackets}	0.25	1
{col_shirts,jackets}	{brown_boots}	0.25	0.5
{brown_boots,jackets}	{col_shirts}	0.25	1

Fonte: MEO, R. et al, 1996.

As cláusulas associadas ao operador MINE RULE permitem definir certas restrições ou condições, tais como considerar um suporte ou confiança mínimos. Para além de ser pssível estabelecer estas restrições, também é possível produzir regras de associação agrupadas em *clusters* (definição de um *cluster* por data, por exemplo) ou produzir regras que obedeçam a determinada hierarquia. A figura abaixo mostra a inclusão da cláusula CLUSTER produzindo, assim, regras sub-agrupadas por data de transacção.

```
MINE RULE ClusteredByDate AS
SELECT DISTINCT 1..n item AS BODY,
   1..n item AS HEAD, SUPPORT, CONFIDENCE
FROM Purchase
GROUP BY customer
CLUSTER BY date
EXTRACTING RULES WITH SUPPORT: 0.01,
                      CONFIDENCE: 0.2
```

Fonte: MEO, R. et al, 1996.

Figura 15 Exemplo de um agrupamento de regras.

O recurso às hierarquias de itens tem por objectivo refinar a descoberta de regras de associação genéricas, obtendo-se assim um número de regras mais reduzido. Suponha-se, por exemplo, que há interesse em extrair regras de associação genéricas que possuam subclasses de botas nos antecedentes e subclasses de calças nos consequentes. Este tipo de problema implica uma especificação de hierarquias de itens, também possível com o operador MINE RULE (figura 16).

```
MINE RULE GeneralizedBootsPantsRules AS
SELECT DISTINCT ancestor AS BODY,
  1..n ancestor AS HEAD, SUPPORT, CONFIDENCE
WHERE  HEAD.ancestor IN (SELECT node
  FROM ItemHierarchy WHERE ancestor ='pants')
  AND BODY.ancestor IN (SELECT node
  FROM ItemHierarchy WHERE ancestor ='boots')
FROM (SELECT  * FROM Purchase, ItemHierarchy
     WHERE node=item)
GROUP BY transaction
EXTRACTING RULES WITH SUPPORT: 0.3,
                      CONFIDENCE: 0.5
```

Fonte: MEO, R. et al, 1996.

Figura 16 Exemplo de uma hierarquia de regras.

O estudo engloba ainda uma referência completa à semântica do operador MINE RULE.

Uma outra proposta de linguagem de consulta a base de dados foi desenvolvida em 1996. Han, em conjunto com mais três investigadores desenvolveu a linguagem DMQL: *Data Mining Query Language* (HAN, J. et al., 1996). Esta linguagem permite extrair conhecimento de bases de dados relacionais, não só sob a forma de regras de associação como também sob a forma de outros modelos de *data mining* relacionados, de alguma forma, com regras (regras de classificação, regras discriminantes, regras de caracterização, etc). À semelhança do operador RULE MINE, também a linguagem DQML tem por base a linguagem SQL. Esta linguagem foi implementada num sistema mais amplo de *data mining*, o DBMiner cujo objectivo principal é o de promover a extracção interactiva de conhecimento, partindo de bases de dados relacionais. Para o desenvolvimento de uma linguagem para extracção de conhecimento, estes investigadores propuseram um conjunto de cinco premissas e considerações, as quais merecem alguma atenção:

- qualquer subconjunto de dados relevante para a tarefa de *data mining* deve ser especificado através da linguagem DMQL, o que significa deixar liberdade ao utilizador para seleccionar um subconjunto de dados, de entre o conjunto de dados de partida, que mais lhe interessa;
- o tipo de conhecimento a descobrir deverá ser especificado pelo utilizador, possibilitando escolher o tipo de conhecimento que melhor se adapta ao problema em análise;
- o conhecimento pré-adquirido deverá sempre ser considerado e disponível para ser aplicado;
- os resultados da extracção de conhecimento devem poder ser expressos sob formas genéricas (conceptuais);
- os utilizadores devem poder definir quaisquer tipos de limites ou restrições, por forma a filtrar conhecimento menos interessante.

Em termos de sintaxe, houve uma preocupação em manter semelhanças com a linguagem SQL, de maneira a que a aprendizagem por parte do utilizador fosse facilitada.

```
DMQLi ::=
use database {database name}
{use hierarchy {hierarchy name} for {attribute}}
{rule spec}
related to {attr or agg list}
from hrelation(s)i
[where {condition}]
[order by {order list}]
{with [{kinds of}] threshold = {threshold value} [for {attribute(s)}]}
```

Fonte: HAN, J. et al., 1996.

Figura 17 Sintaxe da linguagem DQML.

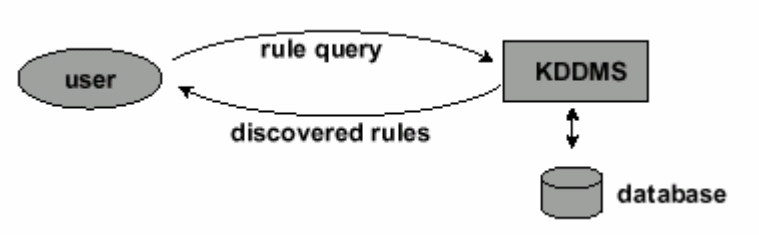
Na figura anterior, está definida a sintaxe genérica da linguagem DMQL. Os caracteres [] representam zero ou uma ocorrência e os caracteres {} significam uma ou mais ocorrências. A instrução {rule spec} define o modelo de regras a utilizar. No caso das regras de associação, deveria utilizar-se a expressão {rule spec} ::= find association rules. Han previu também a definição de hierarquias entre itens, garantindo assim uma forma mais genérica de expressar os resultados da extracção de conhecimento. Uma das possibilidades de aplicação da DMQL defendida por estes investigadores, seria a de funcionar como base para um interface do tipo *Graphical User Interface*[6] (GUI), mais amigável e mais apelativo para o utilizador.

Uma outra proposta de linguagem, baseada em SQL, foi proposta por Morzy e Zakrzewicz, em 1997 (MORZY, T. e M. ZAKRZEWICZ, 1997). A proposta apresenta a MineSQL como uma linguagem declarativa, com objectivos múltiplos, baseada em SQL e que possibilita uma extracção de conhecimento de bases de dados relacionais, de uma forma iterativa e interactiva.

[6] *Graphical User Interface* (GUI): define um ambiente gráfico, através do qual o sistema informático se apresenta e interage com o utilizador.

A linguagem MineSQL apresenta as seguintes características:

- o utilizador especifica uma "*Rule Query*" que representa um conjunto de condições que as regras produzidas devem respeitar;
- o sistema de gestão de todo o processo de KDD (KDDMS) produz as regras de associação que verificam as condições especificadas pelo utilizador;
- as regras de associação são produzidas pelo KDDMS e o utilizador pode avaliar o resultado e, se entender necessário, modificar as condições iniciais e refazer o mesmo processo para obter novo conjunto de regras.



Fonte: MORZY, T. e M. ZAKRZEWICZ, 1997.

Figura 18 Representação do processo interactivo de descoberta de conhecimento.

Morzy e Zakrewicz fazem uma comparação entre a MineSQL para a etapa de *data mining* e a linguagem SQL para as bases de dados tradicionais. Estes investigadores definem a linguagem MineSQL como um interface entre a aplicação cliente e o sistema de *data mining*, ou seja, a aplicação cliente pode estar separada do algoritmo de *data mining* que estiver a ser utilizado. Qualquer modificação realizada no algoritmo não teria, assim, influência nas aplicações entretanto desenvolvidas. A sintaxe da linguagem MineSQL é também semelhante à da SQL: as consultas de *data mining* podem ser combinadas com consultas de SQL, isto é, pode extrair-se conhecimento de resultados SQL e podem ser feitas consultas SQL de resultados da extracção de conhecimento.

Esta linguagem define um novo conjunto de tipos de dados SQL, que são utilizados para armazenar e manipular regras de associação e conjuntos de itens. Na figura seguinte apresenta-se um exemplo de resultado de uma consulta MineSQL.

```
MINE rule, support(rule) AS s., confidence(rule) AS c.
FOR product
FROM shoppings
WHERE 'product=''crescent''' IN body(rule)
AND bodylen(rule) = 2
AND support(rule) > 0.2
GROUP BY trans_id
```

```
rule                                                  s.  c.
----------------------------------------------------- --- ---
product='roll' & product='crescent'->product='pork'  0.4 1.0
product='crescent' & product='pork'->product='roll'  0.4 0.7
```

Figura 19 Consulta em MineSQL e o respectivo resultado (regras de associação).

Os resultados das consultas MineSQL (figura 19) são passíveis de serem colocados numa tabela de dados, como se pode observar pelo exemplo seguinte (figura 20).

```
CREATE TABLE my_rules
(r RULE,
 description CHAR(20))
```

```
INSERT INTO my_rules (r, description)
MINE rule, 'first example'
FOR product, customer
TO time(hour) AS time
FROM shoppings
WHERE head(rule) = 'time=''morning'''
  AND support(rule) > 0.1
```

Fonte: MORZY, T. e M. ZAKRZEWICZ, 1997.

Figura 20 Criação de uma tabela, e posterior inserção do resultado de uma instrução MineSQL.

Os estudos sobre a linguagem MineSQL continuam a ser desenvolvidos, até aos dias de hoje, por Morzy e Zakrewicz (ZAKRZEWICZ, M. 2000 e MORZY, T. 2000).

Em 1998, Goethals, em conjunto com outros dois investigadores, desenvolveu uma técnica que consistia na representação do modelo de *data mining* em base de dados

(neste estudo, o modelo seguido foi o das regras de associação), assim como dos próprios dados utilizados no modelo (GOETHALS, B. et al. 1998). A ideia defendida por estes investigadores é a de que muitos aspectos da interpretação dos resultados de *data mining* (como é o caso da interpretação das regras de associação) podem perfeitamente ser efectuados por consultas (*queries*) colocadas à base de dados, recorrendo a uma linguagem universal (como a *Structured Query Language*). Foi proposta uma estrutura, simples e natural, para representação das regras de associação que pode ser definida da seguinte forma:

- tabela dos conjuntos de itens: representa todos os conjuntos de itens que podem ser encontrados nas regras de associação. É uma tabela constituída por dois campos em que o primeiro identifica o conjunto de itens, e o segundo identifica o item dentro desse conjunto;
- tabela dos suportes: representa o suporte para cada conjunto de itens. É uma tabela constituída por dois campos (um campo identifica o conjunto de itens e o outro o respectivo suporte);
- tabela das regras: as regras de associação são armazenadas numa tabela com cinco campos: o primeiro campo identifica a regra; o segundo o conjunto de itens que representa o antecedente da regra; o terceiro o conjunto de itens que representa o consequente da regra; o quarto e o quinto representam, respectivamente, o suporte e a confiança.

Para além das tabelas acima mencionadas, Goethals sugere que os dados das transacções sejam também colocados na mesma base de dados, utilizando uma tabela constituída por dois campos, identificando, respectivamente, a transacção e o item transaccionado. Segundo este estudo, uma representação deste género permite efectuar diversas operações de pós-processamento, desenvolvidas por inúmeros investigadores, como por exemplo a aplicação de *templates* (KLEMETTINEN, M. et al*,* 1994) ou a utilização de *rule cover*s (TOIVONEN et al., 1995). O objectivo principal deste estudo consistia em procurar demonstrar que uma base de dados relacional, recorrendo às consultas em SQL, oferece uma plataforma poderosa para a implementação de uma ferramenta de pós-processamento para *data mining*. Este estudo considera que o recurso

a uma base de dados orientada para objectos (*OODB*) é uma forte alternativa à base de dados relacional, uma vez que permite representar mais facilmente, e de uma forma mais natural, as consultas à mesma. As linguagens utilizadas nas consultas efectuadas sobre uma *OODB* designam-se por *Object Query Language* ou *OQL* (ver figura 21).

```
select I1.sid, I2.sid
from Itemsets I1, Itemsets I2
where not exists
  (select I3.item
   from Itemsets I3
   where I3.sid=I1.sid and I3.item not in
     (select I4.item
      from Itemsets I4
      where I4.sid=I2.sid))
```

```
select s1, s2
from s1 in Itemsets, s2 in Itemsets
where s1.elements <= s2.elements
```

Fonte: extraída de GOETHALS, B. et al. 1998.

Figura 21 Consulta em SQL (esquerda) e a consulta correspondente em OQL (direita).

A filtragem de regras de associação, recorrendo aos *templates* e às *rule covers*, sob uma forma de consulta à base de dados, foi claramente defendida por Goethals.

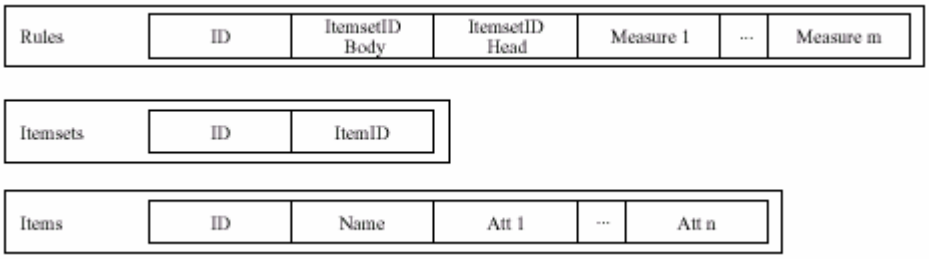
Mais recentemente, quatro investigadores propuseram uma nova abordagem à utilização de bases de dados na descoberta de regras de associação e que se designou por *Rule Cache* (HIPP, J. et al., 2002). Nesta proposta, a principal motivação do estudo baseava-se na ideia de trabalhar com um conjunto, o mais abrangente possível, de regras previamente produzidas por um algoritmo de geração de regras de associação. As regras produzidas seriam armazenadas numa base de dados relacional e todo o trabalho de obtenção de conhecimento seria feito, posteriormente, através de consultas a essa base. Esta abordagem implicava a resolução de duas questões fundamentais.

Em primeiro lugar, havia que dotar o analista de ferramentas capazes de aceder ao *Rule Cache*. A sugestão dos investigadores evoluiu no sentido de se recorrer a uma linguagem de consulta de bases de dados de *data mining*.

Em segundo lugar, o sistema deveria permitir lidar com determinadas restrições ou condições, que não tivessem sido consideradas na fase de obtenção de regras de

associação. Esta questão permitiria generalizar ou especificar determinados casos, assim como focar a observação de certos grupos de regras. A filosofia que sustenta esta abordagem é a de que não é correcto utilizar as restrições como forma de acelerar o processo de produção de regras de associação. Pelo contrário, este estudo defende que é preferível ocupar bastante tempo na produção do maior número de regras possível, sendo que, se pouparia imenso tempo na fase posterior de análise das mesmas.

Estes investigadores propõem que se armazene todas as regras produzidas numa base de dados e se dote o analista de ferramentas de exploração e acesso às regras de associação armazenadas. Na figura abaixo pode observar-se a estrutura relacional das tabelas utilizadas pelo *Rule Cache*. Esta estrutura pode armazenar, para além do suporte e da confiança, normalmente utilizadas na avaliação das regras de associação, outras medidas que se considere fundamentais, assim como certos atributos importantes para a caracterização de determinados itens (figura 22).



Fonte: HIPP, J. et al., 2002.

Figura 22 Representação do modelo de bases de dados que suporta o *Rule Cache*.

Os investigadores defendiam que a exploração das regras deveria ser tão flexível quanto possível, por forma a ser útil a diferentes utilizadores ou a diferentes cenários de extracção de conhecimento. A exploração das regras poderia ser feita recorrendo-se a uma linguagem de consulta de bases de dados específicas para *data mining*. As linguagens DMQL (HAN, J. et al., 1996) e MINE RULE (MEO, R. et al, 1996) são algumas hipóteses sugeridas. No entanto, aqueles investigadores acabaram por criar

uma linguagem de consulta própria, que se aproxima mais das suas ideias de flexibilidade e simplicidade.

```
SelectRulesFrom rulecache                                    (Query 1)
Where conf > 0.75 and lift >= 10;

SelectRulesFrom rulecache                                    (Query 2)
Where 'Airbag' in head and conf > 0.75 and lift >= 10;
```

Fonte: HIPP, J. et al., 2002.

Figura 23 Exemplo de duas consultas à base de dados, utilizando a linguagem do *Rule Cache*.

Estes investigadores desenvolveram ainda um protótipo, a que deram o nome de SMARTSKIP, que implementava as ideias propostas, ou seja, utilizando um algoritmo de geração de regras de associação produzia um conjunto de regras que iriam ser armazenadas numa base de dados, para posterior análise, através de consultas na linguagem proposta.

3 Modelos de regras de associação

Num modelo de regras de associação estão definidas não só as regras produzidas por um determinado algoritmo como também identificadas as diversas características adicionais que possam ser úteis à interpretação das regras, tais como valores de suporte (quer para as regras quer para os conjuntos de itens) valores de confiança, caracterização de todos os itens que intervêm no modelo, entre outras.

Os diversos algoritmos desenvolvidos nestes últimos anos, através da sua implementação em aplicações comerciais têm produzido, para os mesmos problemas, modelos de regras de associação em diferentes formatos. Existem, por exemplo, aplicações que produzem os modelos de regras de associação em formato de texto (*Magnum Opus*, *Weka*, *PolyAnalist*). No entanto, embora seja um formato legível em

qualquer processador de texto, a sua posterior manipulação torna-se algo difícil uma vez que as aplicações não utilizam uma estrutura idêntica.

3.1 Da linguagem XML à linguagem PMML

Durante os últimos anos, diversos modelos de *data mining* foram utilizados na obtenção de conhecimento de bases de dados. No entanto, sempre existiu uma certa dificuldade em representar os resultados obtidos de uma forma que se permitisse reavaliar os diversos modelos produzidos ou até comparar diferentes modelos. Contudo, recentemente, surgiu uma nova forma de exprimir os diversos modelos de *data mining* (entre eles os de regras de associação) gerados pelas inúmeras aplicações informáticas. Designa-se por *Predictive Model Markup Language* ou PMML (*Data Mining Group*) e trata-se de uma linguagem baseada na linguagem XML (*eXtensible Markup Language*). Um documento definido em XML possui um aspecto semelhante a um documento escrito em HTML (*HyperText Markup Language*), popularmente utilizado por milhões de pessoas na Internet. As semelhanças rapidamente se esgotam na simples aparência. Ao contrário da HTML, a XML é uma linguagem extensível, o que significa que, apesar de ser uma linguagem que segue normas rígidas ao nível da sintaxe, deixa total liberdade de criação e definição de elementos (tanto ao nível da designação como do tipo de elemento). Por exemplo, em HTML, para indicar que uma certa parte de um texto faz parte de uma hierarquia, poderiam ser utilizados instruções ou marcadores de tamanho de letra (por exemplo, <font size=”3”>). Estes marcadores, apesar de atribuírem um determinado formato a uma parte de um texto, nada descrevem acerca do conteúdo do mesmo. Em XML, no entanto, é possível criar elementos ou marcadores, que simultaneamente permitam caracterizar quer o formato quer o conteúdo de uma parte de um texto, como por exemplo <resumo> ou <prologo>. Um documento XML pode também conter hierarquias de elementos, devidamente caracterizados, o que permite utilizar um documento XML como um repositório de informação estruturado. No exemplo seguinte, pode-se visualizar um documento XML simples, que poderia ser utilizado para descrever uma mensagem de correio electrónico.

```
<?xml version="1.0" ?>
<email>
       <cabecalho>
              <para>Maria Rosa</para>
              <de>Rui Pedro</de>
              <assunto>Novidades</assunto>
   </cabecalho>
   <corpo>
      Que novidades tens para mim?
   </corpo>
</email>
```

Figura 24 Documento escrito em XML, que descreve uma mensagem de correio electrónico.

Os documentos XML devem seguir um conjunto de normas quer quanto à hierarquia quer quanto ao tipo de dados dos elementos, atributos e entidades a utilizar. Estas normas são definidas através de um conjunto de instruções expressas no próprio documento XML ou num documento separado. Este conjunto de instruções designa-se por *Document Type Definition* (DTD). Tanto um documento XML como o respectivo DTD são escritos em formato de texto normal, pelo que a sua interpretação por diferentes sistemas fica simplificada. As características da linguagem XML, acima descritas, foram suficientes para considerá-la ideal para a criação de uma nova forma de exprimir os diferentes modelos de regras de associação. Assim surge a PMML, uma linguagem XML, definida através de um conjunto de DTD's que exprimem diversos métodos utilizados em *data mining* (regras de associação, árvores de decisão, modelos de regressão, entre outros).

3.2 Representação de modelos em PMML

Tal como acontece com a linguagem XML, também um documento PMML é expresso em formato texto, passível, por isso, de ser interpretado e analisado por diferentes aplicações e plataformas. Um documento PMML permite definir as entidades, os atributos e a estrutura que representam um determinado modelo de *data mining*, independentemente de se tratar de um modelo de regras de associação, árvores de classificação, regressão, rede neuronal ou outro. A flexibilidade desta poderosa

linguagem estende-se, inclusivamente, à representação de novos modelos, sendo que as regras de estruturação de um documento PMML são definidas através do respectivo *Document Type Definition* (DTD). Assim, cada modelo, representado em PMML, segue determinadas regras expressas por um DTD específico e universalmente aceite.

Os pressupostos que levaram à formulação de uma linguagem para a representação de modelos de *data mining* foram estabelecidos da seguinte forma:

- ser universal, isto é, poder ser interpretada por diferentes aplicações, independentemente de quem desenvolve as aplicações ou para que plataforma são desenvolvidas (*Linux*, *Windows*, *Mac*) ;
- ser extensível, permitindo que novos modelos possam ser criados;
- ser portátil, ou seja, facilmente transportada de computador para computador, sem necessidade de instalação de qualquer software;
- ser humanamente compreensível, permitindo conhecer a estrutura do modelo, mesmo sem o recurso a uma aplicação de *data mining*.

Dos pressupostos apresentados, rapidamente se depreende que uma linguagem baseada em XML facilmente cumpre os requisitos, deixando a preocupação com a manipulação ou armazenamento de modelos de *data mining* a cargo das diversas aplicações informáticas, existentes no mercado ou entretanto desenvolvidas. No documento seguinte, observa-se um modelo de regras de associação devidamente especificado, que poderia ter sido gerado por uma aplicação de *data mining* e que está pronto a ser utilizado por qualquer outra aplicação.

```
<?xml version="1.0" ?>
<PMML version="1.1" >

<Header copyright="www.dmg.org" description="exemplo de modelo para R.A.
"/>

<AssociationModel>
<AssocInputStats numberOfTransactions="4" numberOfItens="3"
minimumSupport="0.6" minimumConfidence="0.5" numberOfItensets="3"
umberOfRules="2"/>

<!—existem três itens nos dados de input -->
<AssocItem id="1" value="Cracker" />
<AssocItem id="2" value="Coke" />
<AssocItem id="3" value="Water" />

<!--  e dois registos frequentes com um único item -->
       <AssocItemset id="1" support="1.0" numberOfItens="1">
       <AssocItemRef itemRef="1" />
</AssocItemset>
       <AssocItemset id="2" support="1.0" numberOfItens="1">
       <AssocItemRef itemRef="3" />
</AssocItemset>

<!-- e um registo frequente com dois itens -->
<AssocItemset id="3" support="1.0" numberOfItens="2">
       <AssocItemRef itemRef="1" />
       <AssocItemRef itemRef="3" />
</AssocItemset>

<!-- duas regras satisfazem os requisitos -->
<AssocRule support="1.0" confidence="1.0" antecedent="1" consequent="2" />
<AssocRule support="1.0" confidence="1.0" antecedent="2" consequent="1" />
</AssociationModel>
</PMML>
```

Figura 25 Representação de um modelo de regras de associação em formato PMML.

As vantagens da interpretação e utilização de modelos de regras de associação obtidos a partir de documentos PMML são claras: a existência de uma linguagem única que permite definir diferentes modelos de *data mining* e a sua utilização pelas diferentes aplicações desta área. Prevê-se que, no futuro, as maiores aplicações na área de *data mining* exportem e importem os seus modelos de regras (como de outros métodos) para o formato PMML.

Capítulo II Operadores de regras de associação

Como é possível observar pelo capítulo anterior, foram já vários os estudos desenvolvidos por diversos investigadores, quer no domínio da descoberta de regras de associação (pré-processamento de regras de associação), quer na área da interpretação das regras obtidas pelos respectivos algoritmos (pós-processamento de regras de associação). Os contributos nesse sentido foram surgindo: métodos de agrupamento e resumo de regras, definição de medidas de interesse e regras interessantes e métodos de visualização gráfica. Independentemente do método utilizado em cada estudo, a principal conclusão, que parece reunir algum consenso, aponta no sentido de não ser praticável implementar um método que, por si só, seja suficiente para resolver o problema da análise de uma enorme quantidade de regras obtidas pelos algoritmos de descoberta de regras de associação. O utilizador das ferramentas de descoberta de regras de associação continua a ter a principal decisão sobre que regras de associação melhor servem os seus interesses. Esta conclusão parece ser mais evidente se pensarmos que existe uma grande diversidade de áreas de aplicação da descoberta de regras de associação, analisada pelos mais diferentes tipos de utilizador, com objectivos e interesses bem distintos.

Sendo que o papel do utilizador das ferramentas de *data mining* se considera, cada vez mais, fundamental, este capítulo procura ir de encontro a essa mesma premissa, apresentando uma nova metodologia de ajuda à interpretação de regras de associação. Esta metodologia baseia-se na utilização de um conjunto de operadores que permitem transformar conjuntos de regras do espaço de regras em outros conjuntos de regras, permitindo, desta forma, que o utilizador explore e analise vários conjuntos de regras, um de cada vez. Desta forma, apresentando-se perante um espaço de regras de associação cujas regras se encontram, de alguma forma, estruturadas e relacionadas, torna-se mais fácil explorar grandes conjuntos de regras.

Este método não dispensa a intervenção do utilizador. Pelo contrário, trata-se de um método que exige uma forte intervenção humana, preferivelmente de um utilizador conhecedor do âmbito do problema em análise. No final deste capítulo será ainda

abordada a importância do ambiente Internet para este método, quer para a implementação de um sistema informático quer para um certo paralelismo existente entre a navegação neste ambiente e a navegação pelo espaço de regras de associação.

1 Espaço de regras de associação

As regras de associação, produzidas por um algoritmo de regras de associação para um determinado problema, podem estar, de alguma forma, relacionadas entre si (antecedentes com itens comuns, por exemplo). É possível particionar esse espaço de regras em subgrupos de regras que possuam determinadas características comuns (por exemplo, o antecedentes constituídos pelos mesmos itens). A metodologia apresentada nesta dissertação pretende constituir-se como uma forma de navegação sobre o espaço de regras de associação, permitindo visualizar os diferentes subgrupos de regras, um de cada vez.

Considerando-se $I = \{i_1, i_2, \ldots, i_n\}$ como um conjunto de objectos denominados itens, que assumem valores binários 0 ou 1 (falso ou verdadeiro), considera-se que o espaço de itens I pode ser estruturado, hierarquicamente, segundo uma relação $\subseteq$ entre os conjuntos de itens, que possam existir nas regras de associação em análise. No fundo da hierarquia, considera-se o conjunto de itens vazio $\varnothing$ e, no topo, o conjunto de todos os itens existentes. Esta relação $\subseteq$ corresponde a uma relação de generalidade entre os conjuntos de itens, se estes forem vistos como um conjunto de condições a satisfazer: $A \subseteq B$ significa que A é mais genérico do que B, ou seja, o conjunto de transacções que satisfazem A é maior do que (ou inclui) o conjunto que satisfaz B.

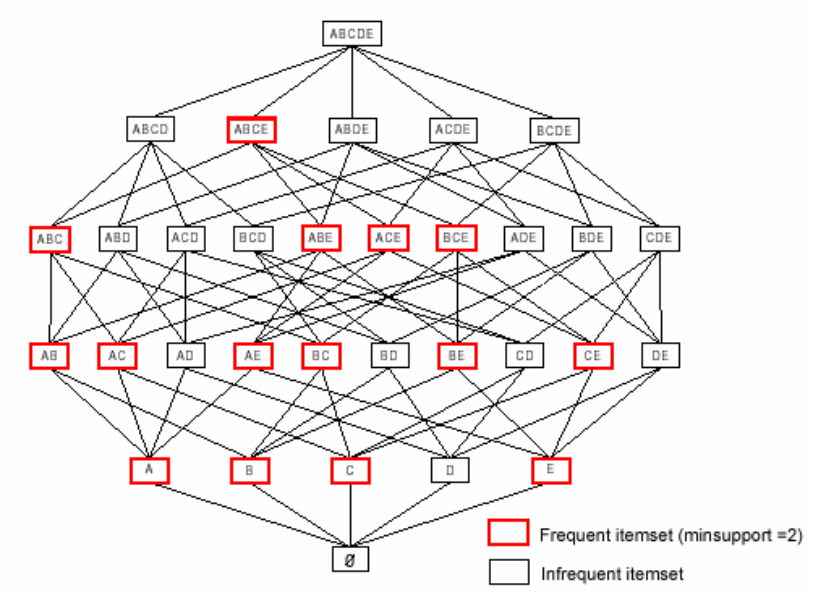
Tome-se, a título de exemplo, uma base de dados de transações representada na tabela seguinte:

Tabela 4 Base de dados de transacções e respectivos itens.

Nº transacção	Itens (produtos)
1	A, C, D
2	B, C, E
3	A, B, C, E
4	B, E
5	A, B, C, E

Fonte: PASQUIER, N. et al., 1998.

Os itens envolvidos nas transacções, originariam, no limite, cerca de 32 conjuntos diferentes, uma vez que existem cinco itens distintos (A, B, C, D e E). Desses 32 conjuntos possíveis, os algoritmos de descoberta de regras de associação (como o *Apriori*) vão identificar os conjuntos frequentes. Finalmente, esses conjuntos frequentes vão produzir as regras de associação.



Fonte: PASQUIER, N. et al., 1998.

Figura 26 Representação do espaço de conjuntos de itens (frequentes e não frequentes).

A figura anterior (figura 26) representa o espaço de todos os conjuntos de itens, e relações entre eles, possíveis de obter a partir da base de dados de transacções anteriormente descrita. Os conjuntos mais frequentes formam, também, um espaço (mais restrito) de conjuntos de itens (PASQUIER, N. et al., 1998). De modo semelhante pode definir-se o espaço de regras de associação.

Para definir e estruturar um conjunto de regras, é necessário relacionar o conjunto de itens frequentes que formam os antecedentes ou os consequentes das regras. Por exemplo, a regra {b, c}→ {e} relaciona-se com regras que possuam o antecedente {b, c}, estruturadas segundo uma relação de especificidade sobre os respectivos

consequentes, mas também se pode relacionar com regras que possuam o consequente {e}, estruturadas segundo uma relação de generalidade sobre os respectivos antecedentes (JORGE et al., 2002a).

Este tipo relação entre regras de associação constitui o espaço de regras, que permite uma navegação entre conjuntos de regras de associação do mesmo espaço. A metodologia dos operadores permite utilizar essa relação por forma a orientar a análise do utilizador em determinadas direcções do espaço de regras.

2 Operadores

A metodologia apresentada nesta dissertação pretende servir de base à navegação sobre um espaço de regras de associação. A referida navegação consiste em transformar conjuntos de regras em outros conjuntos de regras, através da utilização de determinados operadores sobre regras de associação.

Os operadores de regras de associação permitem transformar uma regra $R \in \{\text{regras}\}$ num conjunto de regras $CR \in \{\text{conjuntos de regras}\}$ (JORGE et al, 2002a). Estes operadores actuam sobre uma determinada regra, seleccionada pelo utilizador a partir de um conjunto de regras, e permite "transformar" a mesma num outro conjunto de regras. Esta "transformação" significa, na realidade, que o analista pode obter um conjunto de regras com características semelhantes à regra "transformada", permitindo, assim, direccionar a análise para uma determinada região do espaço de regras. Seguidamente, cada um dos operadores será definido formalmente.

Antecedent generalization **(AntG)**

$AntG(A \rightarrow B) = \{A' \rightarrow B \mid A'$ resulta da subtracção de um ou mais itens de $A\}$

Este operador produz regras idênticas à regra sobre a qual foi aplicado. Face à regra corrente ($A \rightarrow B$), as regras produzidas mantêm o consequente inalterado, sendo que os seus antecedentes são sintacticamente mais simples. Esta operação permite a identificação de itens relevantes ou irrelevantes na regra corrente. Se considerarmos, por exemplo, a regra *Estatisticas_Gerais*, *Industria_e_Energia* $\rightarrow$ *Economia_e_Financas*, aplicando o operador *AntG*, pode obter-se uma regra mais genérica *Estatisticas_Gerais* $\rightarrow$ *Economia_e_Financas*.

Antecedent least general generalization **(AntLGG)**

$AntLGG(A \rightarrow B) = \{A' \rightarrow B \mid$ A' resulta da subtracção de um item de $A\}$

O operador *AntLGG* representa uma versão mais restrita do *AntG*. Neste caso, o antecedente de cada uma das regras produzidas apenas difere em um item em relação ao antecedente da regra corrente.

Consequent generalization **(ConsG)**

$ConsG(A \rightarrow B) = \{A \rightarrow B' \mid B'$ resulta da subtracção de um ou mais itens de $B\}$

Este operador tem uma função inversa ao *AntG*, ou seja, origina uma simplificação do consequente. Produz, também, regras idênticas à regra sobre a qual foi aplicado. Face à regra corrente ($A \rightarrow B$), as regras produzidas mantêm o antecedente inalterado mas cada um dos consequentes são sintacticamente mais simples. Esta operação permite, também, a identificação de itens relevantes ou irrelevantes na regra corrente.

Consequent least general generalization **(ConsLGG)**

$ConsLGG(A \rightarrow B) = \{A \rightarrow B' \mid B'$ resulta da subtracção de um item de $B\}$

O operador *ConsLGG* representa uma versão mais restrita do *ConsG*. Também neste caso, existe uma simplificação do consequente, sendo que, o antecedente de cada uma das regras produzidas apenas difere em um item em relação ao antecedente da regra corrente.

Antecedent specialization **(AntS)**

$AntS(A \rightarrow B) = \{A' \rightarrow B \mid A' \supseteq A\}$

Este operador permite obter regras cujos antecedentes sejam mais específicos, ou seja, cujos antecedentes possuam mais itens que a regra corrente. Neste operador, as regras produzidas têm um suporte inferior, uma vez que se procura uma especificação da regra, mas uma confiança mais elevada que a regra corrente.

Antecedent least specific specialization **(AntLSS)**

$AntLSS(A \rightarrow B) = \{A' \rightarrow B \mid A'$ resulta da adição de um item a $A\}$

À semelhança do operador *AntS*, também este procura uma especialização do antecedente. Neste caso, obtêm-se regras mais específicas que a regra corrente procurando-se regras cujos antecedentes possuam mais um item que o antecedente da referida regra.

Consequent specialization **(ConsS)**

$ConsS(A \rightarrow B) = \{A \rightarrow B' \mid B' \supseteq B\}$

O operador *ConsS* produz resultados semelhantes ao *AntS*, ou seja, também aqui se obtêm regras com um suporte inferior, mas uma confiança mais elevada que a regra corrente. Neste operador, procuram-se regras com consequentes que possuem um ou mais itens para além dos que constituem o consequente da regra corrente. Procura-se assim, obter um conjunto de regras mais específicas que a regra referida. Se considerarmos, por exemplo, a regra *Estatisticas_Gerais* → *Economia_e_Financas*, aplicando o operador *ConsS*, pode obter-se a regra mais específica *Estatisticas_Gerais* → *Economia_e_Financas, Agricultura_e_Pescas*.

Consequent least specific specialization **(ConsLSS)**

$ConsLSS(A \rightarrow B) = \{A \rightarrow B' \mid B'$ resulta da adição de um item a $B\}$

Também o operador *ConsLSS* permite obter um conjunto de regras mais específicas que a regra corrente, através de uma especialização do seu consequente. Esta especificação é um pouco mais restrita do que no caso do operador anterior, uma vez que apenas permite diferir em um a diferença entre os itens dos consequentes das regras obtidas e os itens da regra corrente.

Focus on antecedent **(FAnt)**

$FAnt(A \rightarrow B) = \{A \rightarrow C \mid C$ representa um conjunto qualquer de itens$\}$

O operador *FAnt* produz o conjunto das regras que possuem o mesmo antecedente que a regra corrente. As regras obtidas poderão ser mais específicas ou mais genéricas que a regra corrente, uma vez que não se faz qualquer restrição em relação aos consequentes a procurar. Pode-se ainda definir este operador da seguinte forma: $FAnt(R) = AntG(R) \cup AntS(R)$.

Focus on consequent **(FCons)**

$FCons(A \rightarrow B) = \{C \rightarrow B \mid C$ representa um conjunto qualquer de itens$\}$

De forma similar à do operador anterior, *FCons* produz o conjunto das regras que possuem o mesmo consequente que a regra corrente. As regras obtidas poderão ser mais específicas ou mais genéricas que a regra corrente, uma vez que não existem restrições em relação aos antecedentes a procurar. Uma descrição formal deste operador poderá também ser $FCons(R) = ConsG(R) \cup ConsS(R)$.

Os operadores aqui apresentados têm por base simples propriedades matemáticas de conjuntos e possuem uma semântica intuitiva e clara. Por forma a avaliar da aplicabilidade deste novo método a situações reais, foi desenvolvida uma aplicação informática que funciona em ambiente *web*. Esta solução designa-se por PEAR e será apresentada no capítulo seguinte.

3 Dos operadores à Internet

Cada vez mais, a Internet assume-se como o meio privilegiado de comunicação e partilha de informação entre as pessoas. A *world wide web*, através do seu carácter interactivo e multimédia, é a principal responsável pelo crescimento exponencial da utilização da Internet. As grandes empresas nacionais (privadas ou não) começam também a desenvolver as suas intranets[7] como forma de tirar proveito deste modo de difusão e partilha de informação. Actualmente, as empresas de desenvolvimento de aplicações informáticas elegem este meio como o preferencial para o desenvolvimento dos seus produtos. Jornais, Televisão, Cinema, Investigação, Literatura e muito mais pode ser pesquisado e consultado através da Internet. Os conteúdos publicados na *web* estão acessíveis a partir de qualquer computador, em qualquer parte do mundo.

Pelas razões acima referidas, parece evidente ser de todo o interesse integrar e implementar o método dos operadores no ambiente *web*. Por um lado, é um ambiente cada vez mais familiar aos utilizadores das novas tecnologias, permitindo, por isso, diminuir o processo de aprendizagem que um sistema convencional deste tipo necessitaria. Por outro lado, permite uma portabilidade (não necessidade de instalação) e partilha de dados sem paralelo, facilitando, assim, a sua difusão e discussão entre os utilizadores de *data mining* espalhados pelo mundo. Este ambiente permite, nomeadamente, que diferentes utilizadores possam navegar e analisar o mesmo espaço de regras de associação em simultâneo, podendo, inclusivamente, partilhar opiniões acerca do problema em estudo.

3.1 A metáfora do web browsing

A adicionar aos motivos relacionados com as vantagens do ambiente *web*, para a implementação dos operadores de regras de associação, existe um outro, igualmente interessante, que importa referir. Esse motivo está associado à existência de um certo

[7] Intranet: rede interna de computadores de uma organização que utiliza os protocolos utilizados pela Internet;

paralelismo entre a filosofia dos operadores (navegação entre conjuntos de regras de associação) e a filosofia das páginas *web* e dos seus *hyperlinks*[8] (navegação entre páginas *web*).

Considerando que, a um conjunto de regras de associação corresponde uma página *web*, podemos pensar que os operadores funcionam como *hyperlinks*, permitindo assim, a consulta ou navegação entre conjuntos de regras, como se de simples páginas *web* se tratassem. Por conseguinte, o utilizador consegue compreender, mais facilmente, o modo de funcionamento inerente aos operadores de regras de associação, pois conhece quer o ambiente em que o método dos operadores está a ser utilizado, quer a forma como o espaço de regras pode ser navegado.

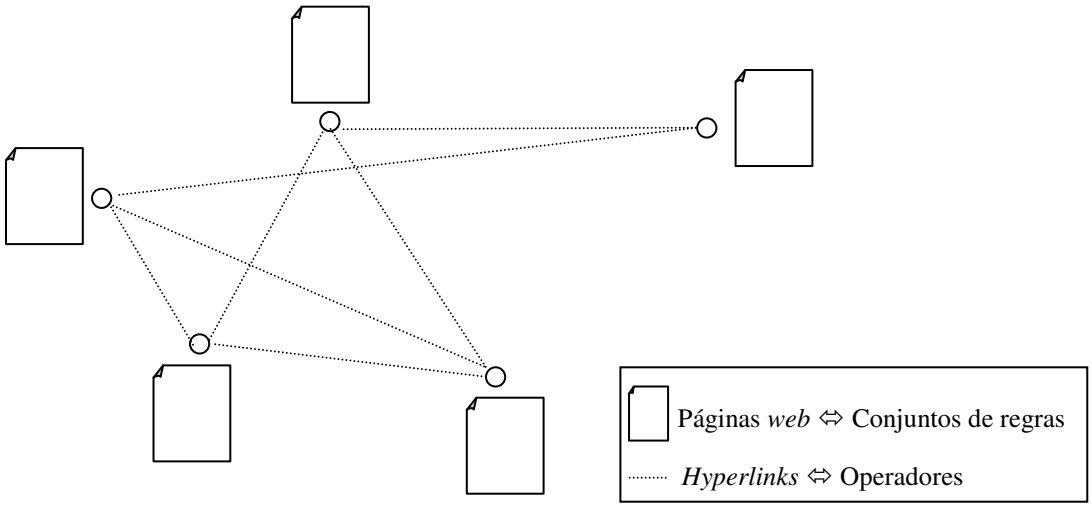


Figura 27 Representação do paralelismo entre conjuntos de regras e páginas *web*.

3.2 A página inicial

As páginas que constituem a *world wide web* encontram-se organizadas por *sites*[9]. Cada um destes *sites* possui uma página inicial, ou seja, um ponto de partida para outras

[8] *Hyperlinks* ou ligações, são o modo de ligação entre diferentes páginas *web* e constituem a principal forma de navegação as mesmas.
[9] *Site* ou sítio: local/endereço da *world wide web* que armazena um conjunto de páginas *web*, normalmente relacionadas através de *hyperlinks*.

páginas (dentro ou fora do *site*). Entre os utilizadores da Internet aquelas páginas são conhecidas por *homepages*. A página inicial de um *site* identifica, de uma forma resumida, a informação que o utilizador poderá encontrar nas restantes páginas do *site*. Trata-se, portanto, de uma página com um objectivo duplo: fornecer um resumo da informação contida nas restantes páginas do *site* e possibilitar ao utilizador o acesso às referidas páginas (através de *hyperlinks*). A informação a colocar nas páginas iniciais é de extrema importância. Se possui demasiada informação pode afastar os utilizadores que se podem sentir perdidos e desorientados. Se possui informação em pouca quantidade pode também levar alguns utilizadores a pensar que não encontram a informação que pretendem nas restantes páginas.

Considerando um certo paralelismo entre *site* e espaço de regras (as páginas *web* que constituem um *site* correspondem aos vários conjuntos de regras que constituem um espaço de regras de associação), torna-se também imperativo que os vários conjuntos de regras produzidos pela aplicação dos diversos operadores possuam também um ponto de partida, ou seja, uma página inicial. Esta página inicial, ou conjunto de regras inicial, permitirá, ao utilizador, iniciar a sua navegação pelo espaço de regras. Tal como acontece na definição das páginas iniciais dos *sites*, também a definição da página inicial para um sistema de operadores de regras de associação pode ser complexa.

Um dos factores de maior peso no sucesso de um sistema que implemente a filosofia dos operadores de regras de associação, senão o de maior peso, é certamente o da definição da "melhor" página inicial para cada espaço de regras. A página com que o utilizador iniciará a sua pesquisa ao espaço das regras será determinante, quer para o desenrolar da análise, quer para a conclusão da mesma. Esta página deverá conter um conjunto de regras suficiente para que o utilizador possa escolher navegar num determinado sentido do espaço de regras. Se este conjunto de regras for demasiado extenso ou demasiado homogéneo pode levar a que o utilizador se sinta confuso e não saiba em que direcção do espaço de regras focar a sua análise.

Para a definição da "melhor" página inicial, poderá fazer algum sentido recorrer a métodos e técnicas amplamente divulgadas e referidas no capítulo anterior. Um

conjunto de regras inicial que resuma (LIU et al., 1999b) o conjunto de todas as regras pode ser considerado um forte candidato a página inicial. De forma semelhante, pode pensar-se que um conjunto de regras com determinadas medidas (BAYARDO, R. e R. Agrawal, 1999) ou que possua as regras mais interessantes (HUSSAIN et al., 1999) do espaço de regras possa ser um óptimo candidato a página inicial.

Capítulo III PEAR, um navegador de regras de associação

Um dos objectivos, desde o início delineados, para esta dissertação era o de aplicar os operadores, teoricamente definidos, sobre conjuntos reais de regras de associação. Concretamente, pretendia-se não só demonstrar a utilização deste método num caso real de interpretação de regras de associação, como também desenvolver mecanismos que permitissem aplicar o método a qualquer outro caso real. Sendo assim, procurou desenvolver-se um sistema que, de uma forma simples, permitisse aferir da utilidade dos operadores, quando aplicados a um espaço de regras de associação produzidos por um qualquer sistema de *data mining*. A designação escolhida para o referido sistema foi PEAR (iniciais de *Post-Processing Environment for Association Rules*). Para a realização de um sistema desta natureza foram levadas em consideração algumas premissas, das quais importa realçar:

- implementação de um interface simples, por forma a reduzir ao mínimo o tempo de aprendizagem do utilizador;
- implementação de um interface independente da plataforma de utilização, ou seja, capaz de funcionar em diferentes tipos de computadores;
- utilização de regras provenientes de um formato o mais universal possível, garantindo uma certa independência em relação aos algoritmos de geração de regras de associação, bem como em relação às aplicações que os suportam;
- utilização dos operadores como forma de navegação no espaço de regras de associação.

Nas páginas seguintes, proceder-se-á a uma descrição mais pormenorizada da forma de utilização, implementação e do modo de funcionamento deste sistema.

1 Utilização

Para a utilização do PEAR, torna-se necessário possuir um PC[10] com ligação à Internet com um *web-browser* instalado ou, em alternativa, um PC com sistema operativo *Windows* com um serviço *web* instalado. O PEAR encontra-se disponível, para utilização, no endereço http://www.3pontos.com/pear. Este sistema apresenta duas funções principais: a leitura de um modelo de regras de associação e a navegação pelo espaço de regras definido pelo respectivo modelo.

A leitura do modelo de regras de associação, é realizada através da opção *Input*, acessível através de uma barra de menu, localizada na parte superior das páginas *web* que constituem este sistema.

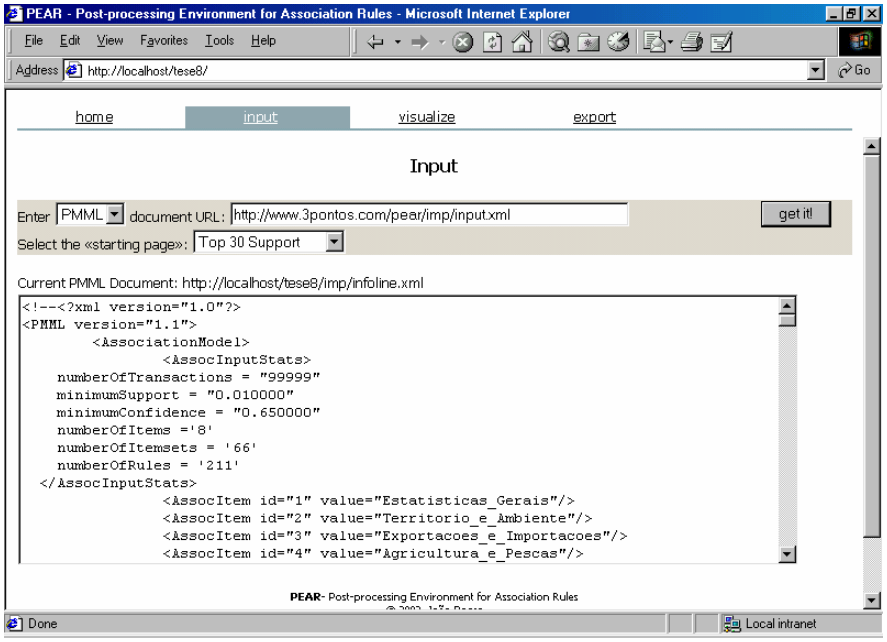


Figura 28 Ecrã do PEAR - leitura de um modelo de regras de associação (opção *Input*).

A figura anterior apresenta um modelo de regras de associação, em formato PMML, carregado e validado através da opção *Input*, disponível na barra de menu (topo superior do ecrã).

[10] PC: sigla internacionalmente utilizada que identifica o Computador Pessoal (*Personal Computer*)

Após a leitura do modelo, o utilizador pode iniciar a análise das regras de associação, definindo qual o conjunto de regras inicial (página inicial) a analisar e seleccionando, posteriormente, a opção *Visualize*, disponível também no menu localizado na parte superior do ecrã.

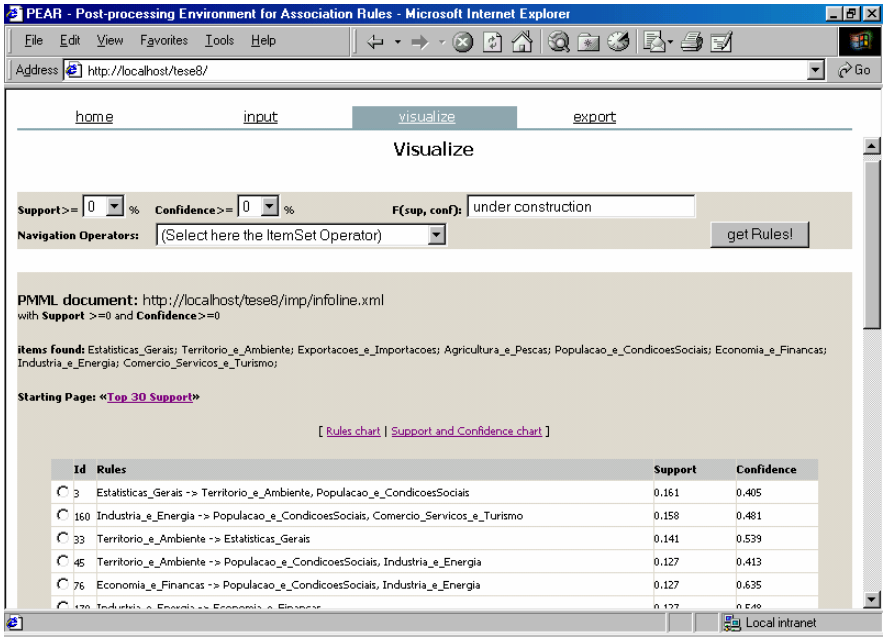


Figura 29 Ecrã do PEAR - visualização de um conjunto de regras inicial (opção *Visualize*).

A figura anterior apresenta um conjunto inicial de regras, a partir da qual o utilizador pode consultar outros conjuntos, recorrendo ao método dos operadores, disponível na área intermédia do ecrã. Após a selecção do operador a utilizar e da regra a "transformar", a visualização de um novo conjunto de regras é iniciada através do botão "get Rules". Este processo de consulta é iteractivo, ou seja, permite uma sucessiva aplicação de operadores aos conjuntos de regras que vão sendo, sucessivamente, visualizados. Para cada conjunto de regras consultado, é possível efectuar uma representação gráfica das mesmas, utilizando para isso as opções "Rules chart" e "Confidence and Support chart" (ver figura 30).

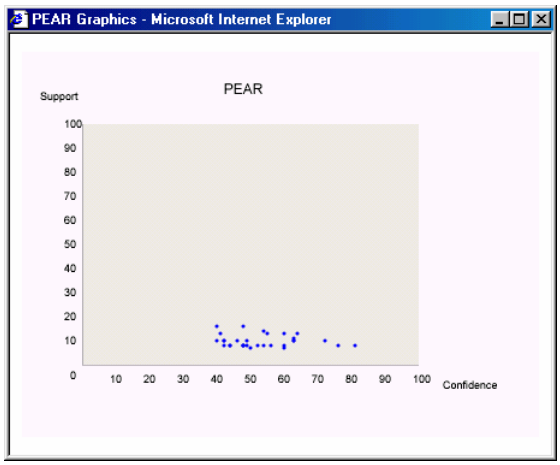


Figura 30 Ecrã do PEAR – componente gráfica (opção *Rules chart*).

A figura anterior representa, graficamente, os valores de confiança e suporte das regras visualizadas sob a forma de tabela na página inicial. Cada regra é representada na figura por um ponto azul, sendo que é possível verificar a que regra corresponde cada ponto, através da deslocação do cursor do rato sobre um determinado ponto. O utilizador pode, através do gráfico, identificar uma regra que tenha uma combinação adequada de suporte e confiança e explorar o espaço de regras a partir dela.

Deste modo, torna-se possível efectuar uma navegação pelo espaço de regras de associação, visualizando um conjunto de regras de cada vez, sob a forma de tabela de regras e sob a forma gráfica.

2 Escolha da plataforma

A escolha da plataforma informática sobre a qual foi implementado este sistema consistiu no sistema operativo *Microsoft Windows NT 4.0* (ou versões superiores). Note-se que, no entanto, o sistema corre noutros sistemas *Windows* (tais como *95*, *98*, *2000* ou *XP*). A opção por este sistema operativo teve que ver com algum conhecimento adquirido ao longo de alguns anos de experiência a trabalhar nesta plataforma, mas, sobretudo, pelas facilidades que a *Microsoft* disponibiliza em termos de linguagem XML. O esforço e o contributo desta empresa em torno da XML, quer em termos de apoio ao desenvolvimento de definição de padrões, quer em termos de criação de ferramentas de suporte à XML tem sido simplesmente notável.

Uma boa alternativa a esta plataforma seria, claramente, a plataforma JAVA. Um sistema como o PEAR, desenvolvido em JAVA garantiria todas as premissas anteriormente abordadas. No entanto, o tempo de aprendizagem desta linguagem seria demasiado longo para os propósitos desta dissertação. Contudo, não fica de lado a possibilidade de, no futuro, o PEAR poder evoluir para a plataforma JAVA.

3 Tecnologias envolvidas

Como já foi anteriormente referido, este sistema foi essencialmente desenvolvido recorrendo a tecnologias *Microsoft*, tendo sempre a preocupação de que fosse um sistema independente da plataforma que o utilizador possuísse. O PEAR pode ser classificado de sistema cliente-servidor, uma vez que depois de instalado num computador servidor pode ser utilizado por diferentes clientes, através da Internet (mais objectivamente da *world wide web*). Para a instalação do servidor, é necessário um computador com sistema operativo *Windows* (versão *98*, *Millenium*, *NT*, *2000* ou *XP*). Como software adicional, é necessário existir um servidor *web* da *Microsoft* em funcionamento assim como um dispositivo de ligação a base de dados *Microsof Access* (ODBC *Driver*). Uma vez que se utiliza uma ligação ODBC, podem utilizar-se outros SGBD. Para a utilização do PEAR, pode ser usada a máquina em que é instalado, ou um computador com acesso, via *http*[11] ao servidor acima referido.

Na implementação e funcionamento do PEAR estão envolvidas as seguintes tecnologias:

- IIS, *Microsoft Internet Information Server* (no *Windows 98* ou *Windows Millenium* pode ser substituído pelo serviço *Personal Web Server*)
- ASP, *Active Server Pages* e VbScript
- SQL, *Structured Query Language*
- JavaScript
- DOM, *Document Object Model*
- SVG, *Scalable Vector Graphics*

Seguidamente, cada uma destas tecnologias será descrita sucintamente e explicar-se-á o modo como foram aplicadas e conjugadas no PEAR.

[11] Http: protocolo do nível de aplicação para sistemas de informação distribuídos, que colaboram entre si e possuem capacidades de hipermédia, que tem sido usado na *World Wide Web* desde 1990.

3.1 Internet Information Server (IIS)

O *Internet Information Server* (IIS), desenvolvido pela *Microsoft*, tem a função de servidor *web*, ou seja, trata-se de um serviço que disponibiliza páginas na Internet (*world wide web*), através de um protocolo *http*. A particularidade deste servidor *web*, quando comparado com outros concorrentes, prende-se com a possibilidade de recorrer a páginas *web* geradas dinamicamente, isto é, criadas no momento em que o utilizador está a efectuar a sua consulta, sem necessidade de instalar componentes adicionais. Esta funcionalidade permite que, por exemplo, diferentes utilizadores possam estar a visualizar a mesma página mas com conteúdos diferentes. Este tipo de páginas dinâmicas será abordado no ponto seguinte. O IIS é fornecido como um complemento gratuito do sistema operativo *Windows* NT, *Windows 2000* ou *Windows XP*. A partir do momento em que se executa um serviço IIS, num dos sistemas operativos anteriormente descritos, e se possui uma ligação à Internet, podem disponibilizar-se páginas na *web* acessíveis a partir de qualquer outro computador, ligado à rede mundial. Uma variante mais simples do IIS, designada por *Personal Web Server* (PWS), é utilizada por sistemas operativos menos poderosos, tais como o *Windows 98* e *Windows Millenium*, podendo, no entanto, também ser utilizada pelo PEAR, em modo local (servidor e cliente no mesmo PC).

3.2 Active Server Pages e VbScript

Os documentos ou páginas *Active Server Pages* (ASP) são páginas *web* processadas, previamente, através de um servidor IIS ou PWS e disponibilizadas, posteriormente, aos utilizadores (computador cliente). Um documento ASP integra comandos HTML e comandos do tipo *script* (linguagem com sintaxe simples e executada sequencialmente). Estes comandos *script* podem ser definidos em linguagem VbScript ou JScript (ambas desenvolvidas pela empresa *Microsoft*). A linguagem JScript, à semelhança da linguagem JavaScript da empresa concorrente *Netscape*, procura seguir as recomendações da *European Computer Manufacturing Association* (ECMA-262

standard[12]). No PEAR, optou-se por utilizar a VbScript, uma vez que as páginas ASP são independentes da plataforma cliente. Quando uma página ASP é invocada pelo utilizador, estes comandos *script* são executados pelo IIS produzindo código HTML normal que é enviado para o computador cliente. Este tipo de páginas, sendo geradas no momento em que o utilizador as invoca e podendo variar o conteúdo que o cliente visualiza, são designadas, por isso, de páginas dinâmicas. As páginas ASP possuem algumas características específicas, que as distinguem das tradicionais páginas estáticas:

- permitem ligações a bases de dados, possibilitando a execução de diversas operações sobre a mesma (consultas, inserções e eliminações de registos) através de comandos em linguagem SQL;
- permitem uma adaptação do conteúdo disponibilizado ao utilizador (cliente) através de uma linguagem simples desenvolvida pela *Microsoft* (VbScript ou JScript);
- permitem obter e processar informação proveniente dos utilizadores e enviada do seu computador para o servidor.

As páginas ASP são utilizadas pelo PEAR para a manipulação da base de dados que suporta toda a informação obtida a partir do modelo inicial de regras de associação (em formato PMML). Nesta manutenção da base de dados, está implícita a execução de todo o tipo de operações, através de comandos SQL.

Este tipo de páginas é útil também para, por exemplo, obter informação do utilizador, sendo desta forma que o PEAR recebe as ordens do utilizador, para navegação pelo espaço de regras. O processo de obtenção de informação proveniente do utilizador, bem como do seu posterior processamento, é efectuado recorrendo à linguagem VbScript. Esta linguagem foi desenvolvida pela Microsoft e apresenta uma sintaxe semelhante ao *Microsoft Visual Basic*. Os documentos ASP podem conter instruções de processamento em VbScript ou JScript.

[12] ECMA: associação internacional fundada em 1961 que se dedica a definir standards de sistemas de informação e comunicação.

No caso do PEAR, optou-se por utilizar VbScript, para instruções processadas pelo servidor, sendo que para instruções processadas pelo computador cliente decidiu-se por JavaScript. Deste modo, torna-se mais simples interpretar os documentos ASP, mantendo uma total independência, em relação ao *web-browser* utilizado pelo computador cliente. A linguagem VbScript é também utilizada para a manipulação dos documentos PMML e documentos SVG (ambos baseados em XML). Para esta manipulação, a VbScript recorre ao *Document Object Model* (DOM), que será abordado seguidamente.

3.3 Base de dados e Structured Query Language (SQL)

Quando se acede a uma base de dados relacional, na realidade está a utilizar-se uma versão padronizada da linguagem *Structured Query Language* (SQL). Os principais sistemas de gestão de base de dados, por exemplo da *Oracle*, da *Microsoft* ou da IBM suportam SQL. Originalmente, a SQL foi definida e implementada pela IBM para um sistema de base de dados designado por SYSTEM R. Rapidamente se compreendeu a importância de uma linguagem deste tipo para as bases de dados, pelo que se desenvolveram outros sistemas que adoptaram algumas variantes da mesma. Devido à popularidade da linguagem, iniciaram-se esforços por criar uma linguagem comum a todos os vendedores, sendo que a última versão da linguagem SQL, designada por SQL-99 ou SQL-3, foi aprovada pelo *American National Santards Institute* (ANSI) em 1999.
A SQL é uma linguagem não orientada para procedimentos, ao contrário das linguagens mais populares, como o *Visual Basic*, C, C++, Java, VbScript ou JavaScript. Numa linguagem orientada para o procedimento as instruções definem *como* processar a informação; na linguagem não orientada para procedimentos as instruções descrevem *que* informação se procura, deixando para o sistema gestor da base de dados o trabalho de definir o melhor método para a obter.

No caso do PEAR, o Sistema de Gestão de Base de Dados (SGBD) utilizado para suporte ao modelo de regras de associação foi o *Access*, também da *Microsoft*. Aliada à simplicidade inerente a este SGBD, pode observar-se alguma perda de ineficácia, nomeadamente na natureza multi-utilizador, como é o caso de uma aplicação para a

Internet. A utilização, pelo PEAR, de um SGBD mais evoluído, como é o caso do SQL Server da *Microsoft*, é perfeitamente possível e as adaptações de programação necessárias são insignificantes. Esta solução, no entanto, implicaria a perda de alguma portabilidade do sistema, uma vez que um SGBD mais evoluído tem repercussões óbvias ao nível do custo de implementação do PEAR e não seria suportado por qualquer entidade que desejasse utilizar o PEAR com acesso a este SGBD.

Independentemente do SGBD utilizado, os comandos de SQL podem ser invocados recorrendo a ferramentas próprias dos sistemas de gestão de bases de dados, mas podem também ser executados por intermédio de uma linguagem de programação como é o caso da linguagem VbScript. No caso do PEAR, à semelhança de qualquer aplicação que funcione com páginas ASP, os comandos de SQL são executados e incorporados na linguagem VbScript.

Na esquema seguinte (figura 31), estão representadas as tabelas assim como a relação existente entre as mesmas, no modelo de dados relacional que suporta o modelo de regras de associação obtido de PMML.

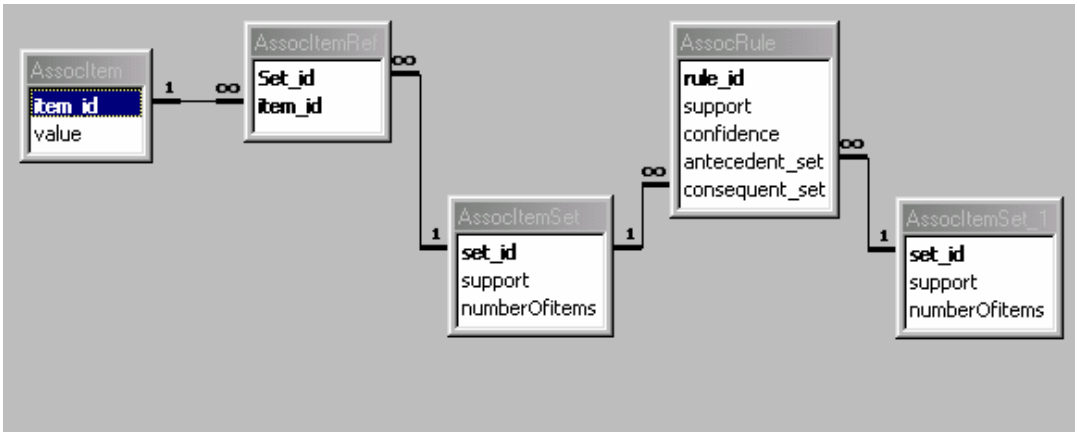


Figura 31 Modelo de dados utilizado pelo PEAR.

3.4 JavaScript

Tal como acontece com a JScript, a linguagem JavaScript também segue as recomendações da *European Computer Manufacturing Association* (ECMA-262 standard). Decidiu-se utilizar a JavaScript como linguagem de processamento de informação no lado do cliente, uma vez que esta linguagem é perfeitamente compatível com os dois *web-browsers* mais utilizados pelos internautas[13] (*Microsoft Internet Explorer* e *Netscape Navigator*). Desta forma, a independência do sistema face à plataforma do utilizador (um dos principais pressupostos do sistema PEAR) fica garantida. Esta linguagem de *script* é bastante útil na validação de dados e na interacção com o utilizador. No caso do PEAR, existem algumas funcionalidades que dependem da JavaScript, tais como o menu principal de opções ou a interacção da parte gráfica com o utilizador. No futuro prevê-se uma maior utilização desta linguagem, quer como forma de reforçar a interacção quer como forma de validar os dados introduzidos pelo utilizador.

3.5 Document Object Model (DOM)

O *Document Object Model* (DOM) define um interface, que permite manipular, seguindo uma estrutura em forma de árvore, documentos XML ou baseados em XML, entre outro tipo de objectos. O DOM está definido como recomendação, pelo *World Wide Web Consortium*, desde Outubro de 1998 (W3C DOM *Level 1 specification*). Acedendo à estrutura interna de cada documento XML, o DOM facilmente o representa sob a forma de árvore, permitindo assim uma manipulação simples de todos os seus elementos. No caso do PEAR, o recurso ao DOM é útil para o processamento inicial do modelo das regras de associação definido em PMML, assim como para a manipulação dos histogramas definidos em documentos *Scalable Vector Graphics* (SVG), uma vez que ambos são documentos baseados em XML. Durante o período de desenvolvimento do PEAR, colocou-se a hipótese de não se utilizar uma base de dados para suporte do

[13] Internauta: denominação atribuída ao utilizador da Internet.

modelo de regras de associação, recorrendo-se apenas ao DOM. A manipulação de uma base de dados, via Internet, processa-se de forma mais lenta que a manipulação de um documento escrito em DOM. No entanto, o DOM não permite efectuar consultas complexas sobre os elementos e nós do documento, pelo que, desde logo prevaleceu a decisão pela utilização da base de dados.

3.6 Scalable Vector Graphics (SVG)

Um documento escrito em *Scalable Vector Graphics* (SVG) é um documento baseado em XML que especifica e define elementos gráficos, que podem ser visualizados por um *web-browser*. Deste modo, o processamento de documentos SVG é semelhante ao processamento de um documento XML. Os gráficos (documentos) em SVG oferecem as vantagens dos gráficos vectoriais adicionando características próprias de uma linguagem XML, permitindo ainda o uso de interactividade. O SVG oferece poderosas capacidades de texto, interactividade (permitindo inclusivé a utilização de *hyperlinks*) e imagens (fixas ou em movimento), tudo no formato vectorial. A linguagem SVG é também uma recomendação do W3C (W3C *Recommendation*, *04 September 2001*) que a definiu como uma linguagem que descreve vectores bidimensionais e simultaneamente os permite combinar com gráficos não vectoriais. O PEAR utiliza o interface DOM, por intermédio do VbScript, para a criação dos gráficos em formato SVG. Para a manipulação dos gráficos, após a sua criação, o PEAR recorre ao JavaScript. Com esta tecnologia, torna-se mais simples desenvolver elementos gráficos com alguma complexidade e, por outro lado, é possível tornar estes elementos interactivos, controlando eventos produzidos pelo rato ou pelo teclado. Para visualizar qualquer "gráfico" no formato SVG, é necessário fazer-se o *download*[14] de um *plug-in*[15] distribuído gratuitamente pela empresa *Adobe Systems Incorporated*.

[14] *Download*: termo inglês, vulgarmente utilizado pelos utilizadores da Internet, que significa a transferência de um ficheiro de um computador remoto (servidor) para o computador cliente.
[15] *Plug-in*: termo inglês, utilizado quando se quer referir a um pequeno programa ou módulo que, depois de instalado no computador, permite dotar uma determinada aplicação (neste caso um navegador de páginas da Internet) de novas funções.

Nas figuras seguintes pode observar-se um exemplo de código de uma página HTML com o "gráfico" SVG incorporado e o respectivo *output*.

```
<html>
<head><title>teste</title></head>

<body bgcolor="white">

<embed src="teste.svg" name="SVGSlide2" width="100%"
height="100%" type="image/svg-xml"
pluginspage="http://www.adobe.com/svg/viewer/install/">
</embed>

</body>
</html>
```

Figura 32 Documento escrito em HTML que incorpora um ficheiro SVG.

```
<?xml version="1.0" ?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG 20000802//EN"
"http://www.w3.org/TR/2000/CR-SVG-20000802/DTD/svg-20000802.dtd">

<svg width="250" height="200">
   <path d="M 30 0 L 100 100" />
       <text x="20" y="80" style="font-size:40; font-weight:400;
font-family:Verdana; font-style:italic; fill:red">Texto</text>
</svg>
```

Figura 33 Código fonte de um documento SVG.

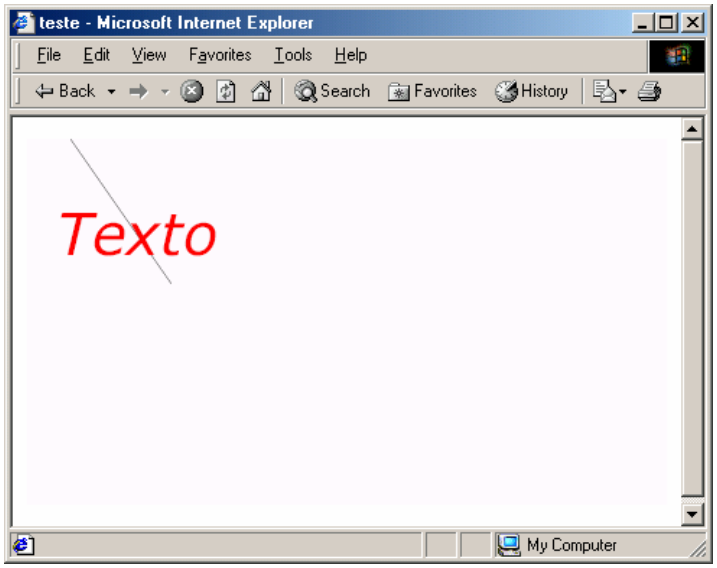


Figura 34 Output produzido pela página HTML com um documento SVG incorporado.

4 Estrutura funcional

O PEAR possui uma forma de funcionamento extremamente intuitiva. Após a entrada de dados, proveniente de um modelo de regras de associação em PMML, as regras (assim como informação sobre os itens e conjuntos de itens que as constituem) são armazenadas numa base de dados, para posterior manipulação. Essa manipulação consiste na navegação pelo espaço de regras de associação representado no modelo considerado. A estrutura funcional do PEAR, em que se pode observar está esquematizada na figura seguinte.

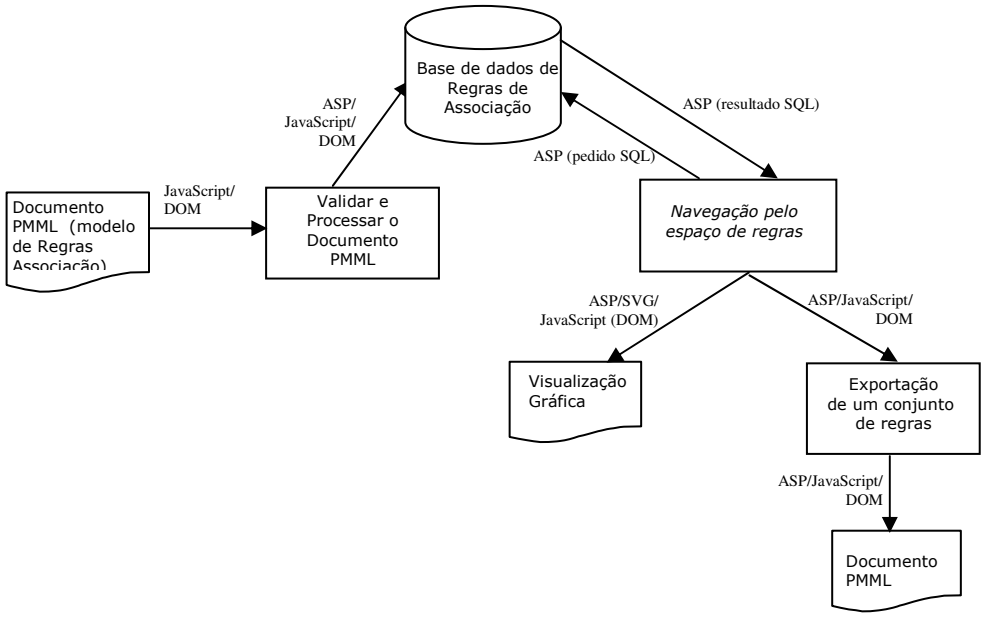


Figura 35 Representação do funcionamento do PEAR.

Em termos gerais, pode considerar-se que o sistema PEAR está organizado em três módulos principais:

- *Input*: leitura de um modelo de regras de associação;
- *Browsing*: navegação no espaço de regras (que permite a visualização, quer tabular quer gráfica, dos diversos conjuntos de regras de associação);

- *Output*: possibilidade de exportar, para um formato universal (PMML), um conjunto de regras definidas pelo utilizador.

O sistema PEAR necessita de um modelo inicial de regras de associação, em formato XML (PMML), gerado por qualquer aplicação informática de *data mining* que produza regras de associação e as exporte para um documento formato PMML. Este documento poderá ser interpretado pelo PEAR, no módulo de *Input*. O documento só será visualizado e posteriormente armazenado na base de dados se for um documento válido, isto é, se obedecer às normas correctas de um documento PMML.

Após a referida validação do documento PMML o sistema procede ao seu tratamento, transformando o respectivo documento no modelo de dados da base de dados em que serão armazenadas as regras de associação. A partir do momento em que as regras de associação, os itens e os conjuntos de itens estão armazenados na base de dados, o sistema está disponível para efectuar a navegação pelo espaço das regras utilizando os operadores de regras.

O módulo de *Output* (módulo que ainda não foi desenvolvido) corresponde à exportação de um determinado grupo de regras de associação para o formato PMML.

5 Algoritmo de navegação

A secção anterior apresenta uma visão geral do modo de funcionamento do conjunto de páginas *web* que constituem o sistema PEAR. Nesta secção, serão abordados alguns detalhes de programação relacionados com a implementação da navegação pelo espaço de regras (módulo de *Browsing*), através da utilização do método dos operadores. Este módulo possui uma certa complexidade uma vez que utiliza recursos complexos de gerir em ambiente *web*, nomeadamente a manipulação de base de dados e a interacção com o utilizador.

O objectivo deste módulo é o de apresentar, mediante as escolhas definidas pelo utilizador, diversos conjuntos de regras de associação provenientes do modelo de regras carregado para uma base de dados. Para além de uma visualização das regras de associação sob um aspecto de tabela ou lista de regras permite, ainda, uma visualização com elementos gráficos que fornece elementos complementares de análise à visualização tabular. O utilizador tem disponíveis diversos elementos que lhe permitem orientar a sua análise para um determinado sentido do espaço de regras. Pode utilizar os operadores de regras de associação, consultando conjuntos de regras com determinadas características. Pode ainda, por exemplo, definir valores mínimos de confiança e de suporte, filtrando assim, segundo estas medidas, as regras que são visualizadas.

A programação deste módulo foi realizada em *Active Server Pages* (ASP), uma vez que há a necessidade de aceder à informação contida na base de dados e de adaptar o conteúdo da página ao conteúdo da informação proveniente da base de dados. A utilização de uma base de dados, apesar de não ser imprescindível para a implementação do método dos operadores, permite obter mais facilmente informação sobre o modelo. Esta facilidade tem que ver com a versatilidade no acesso à informação proporcionada pela linguagem SQL disponível no sistema de gestão da base de dados (neste caso, *Microsoft Access*). Pelo contrário, a não utilização de uma base de dados implicaria o recurso à tecnologia DOM que, devido à sua simplicidade, obriga a um esforço de programação adicional resultando, também, num acesso mais lento.

Na figura seguinte está definido, em Pseudocódigo, o algoritmo utilizado neste módulo.

```
1- Lê Modelo de regras de associação (documento PMML)
2- Se documento não válido ir para Fim
3- Modelo não existe na Base de Dados (BD)?
    3.1- Então colocar elementos que caracterizam o modelo na
       BD
4- Construir Página Inicial
5- Repetir até Evento = terminar
    4.1- Detecta Evento
    4.2- Se Evento = aplicar operador ou Evento = alterar
    suporte ou Evento = alterar confiança
        4.2.1 – Construir a consulta em linguagem SQL, com base
        nas escolhas do utilizador
        4.2.2- Executar a consulta à BD
        4.2.3- Construir a página com o conjunto de regras
        obtidos pela consulta à BD
    4.3- Se Evento = visualizar regras
        4.3.1- Construir gráfico correspondente ao conjunto de
        regras visualizadas
6- Fim
```

Figura 36 Pseudocódigo do módulo de navegação.

Como se observa pelo pseudocódigo apresentado (passo 1.), o módulo de navegação inicia-se pela leitura do documento PMML. Este processo é realizado recorrendo a um objecto do tipo MSXML (*Microsoft XML*). A manipulação deste objecto faz-se, como já foi referido anteriormente, utilizando a tecnologia DOM. Pode observar-se, na figura seguinte, a forma de criação deste objecto, a partir do qual se pode retirar toda a informação relativa ao documento PMML.

```
set pmml =
CreateObject("microsoft.xmldom")
pmml.async=false
pmml.load(endereco)
```

Figura 37 Criação do objecto (MSXML) correspondente ao documento PMML.

A criação do objecto MSXML pode ser ou não executada com sucesso, ou seja, o documento pode ou não estar válido (coerente com as regras definidas para documentos XML e, mais concretamente, para um documento PMML). No caso de não se tratar de um documento válido o sistema alerta o utilizador para a não validade do mesmo. No

caso do documento ser válido, o sistema verifica se o documento já existe na base de dados, sendo que, se já existir, o processo de colocação na referida base da informação proveniente do documento PMML não é executado. Este processo implica a manipulação do documento PMML utilizando a tecnologia DOM. A título de exemplo, apresenta-se, seguidamente, uma rotina que obtém informação sobre os itens existentes no documento (que irão constituir as regras de associação).

```
' ------------------------------------------------------------
' percorre o documento PMML em busca de itens
' ------------------------------------------------------------
function obtem_items(nodulo_i)

   ' --- obter o número de itens
   set nodulo_i=nodulo_i.getElementsByTagName("AssocItem")
   num_items=nodulo_i.length

   ' --- obter informaçao para cada item
   for i=0 to num_items-1
      num_atributos=nodulo_i.item(i).Attributes.length
      for a=0 to num_atributos-1
         if nodulo_i.item(i).Attributes.item(a).nodeName="id" then

   items_array_id(i)=nodulo_i.item(i).Attributes.item(a).nodeValue
         end if
         if nodulo_i.item(i).Attributes.item(a).nodeName="value" then
            items_dsg=items_dsg &
nodulo_i.item(i).Attributes.item(a).nodeValue

   items_array(i)=nodulo_i.item(i).Attributes.item(a).nodeValue
            if i<>num_items-1 then items_dsg = items_dsg & ", "
         end if
      next
   next
end function
```

Figura 38 Função que utiliza a tecnologia DOM para interpretar documento PMML.

Como se pode constatar pela observação da função representada na figura 38, as instruções DOM são simples, permitindo aceder ao documento de uma forma estruturada (estrutura em árvore), estando a informação sobre os itens, os conjuntos de itens e as regras de associação em ramos perfeitamente separados. Após a obtenção de informação sobre os itens, conjuntos de itens e regras, procede-se à colocação dessa informação na base de dados.

A partir deste momento, o sistema tem acesso a informação suficiente para apresentar, ao utilizador, um qualquer conjunto de regras de associação existente no espaço de regras. Cada um destes conjuntos de regras, quer se trate do conjunto inicial de regras (página inicial) quer de um outro conjunto de regras, é obtido a partir de uma consulta à base de dados, realizada através de uma instrução SQL. Esta instrução SQL é criada dinamicamente, ou seja, é criada com base nas escolhas do utilizador para cada conjunto de regras.

```
' operador 10 = focus on consequent
if operador=10 then
   ' número de itens do CONSEQUENT
   SQLQUERY1 = "SELECT * FROM AssocItemSet Where set_id='" & consequent_set &
"'"
   Registo1.open SQLQuery1, conexao

   number_items=registo1("numberofitems")

   ' obter a identificação das regras que possuem o mesmo consequente
   SQLtemp = "SELECT assocrule.rule_id, ae.numberOfitems INTO temp  "
   SQLtemp = SQLtemp & "FROM assocrule INNER JOIN
   Mostra_antecedentes_existentes AS ae ON assocrule.antecedent_set =
   ae.antecedent_set "
   SQLtemp = SQLtemp & "WHERE assocrule.consequent_set='" & consequent_set &
"' "
   SQLtemp = SQLtemp & "AND assocrule.antecedent_set<>'" & antecedent_set &
"'"
   conexao.Execute   SQLtemp
end if


SQLQuery2 = "SELECT  temp.rule_id, ant.item_id, "
SQLQuery2 = SQLQuery2 & "ant.value, ant.Set_id, "
SQLQuery2 = SQLQuery2 & "ant.support,  "
SQLQuery2 = SQLQuery2 & "ant.confidence, "
SQLQuery2 = SQLQuery2 & "ant.type, "
SQLQuery2 = SQLQuery2 & "ant.numberOfitems "
SQLQuery2 = SQLQuery2 & "FROM temp INNER JOIN
Mostra_antecedentes_de_cada_regra  AS ant "
SQLQuery2 = SQLQuery2 & "ON temp.rule_id = ant.rule_id   "
SQLQuery2 = SQLQuery2 & "WHERE ant.support>="   & suporte & " AND
ant.confidence>=" & confianca & " "
SQLQUERY2 = SQLQUERY2 & "ORDER BY temp.rule_id, ant.type "
SQLQuery2 = SQLQuery2 & "UNION "
SQLQuery2 = SQLQuery2 & "SELECT  temp.rule_id, cons.item_id, "
SQLQuery2 = SQLQuery2 & "cons.value, cons.Set_id,  "
SQLQuery2 = SQLQuery2 & "cons.support, "
SQLQuery2 = SQLQuery2 & "cons.confidence, "
SQLQuery2 = SQLQuery2 & "cons.type, "
SQLQuery2 = SQLQuery2 & "cons.numberOfitems "
SQLQuery2 = SQLQuery2 & "FROM temp INNER JOIN
Mostra_consequentes_de_cada_regra  AS cons"
SQLQuery2 = SQLQuery2 & "ON   temp.rule_id = cons.rule_id "
SQLQuery2 = SQLQuery2 & "WHERE cons.support>="  & suporte & " AND
cons.confidence>=" & confianca & " "
SQLQUERY2 = SQLQUERY2 & "ORDER BY temp.rule_id,cons.type "

Registo.open SQLQUERY2, conexao
```

Figura 39 Criação dinâmica da consulta SQL à base de dados (Focus on Consequent).

O exemplo anterior (figura 39) reflete a criação da instrução SQL que permite executar o operador "Focus on Consequent" sobre determinada regra "transformando", desta forma, um determinado conjunto de regras num novo conjunto de regras . Neste caso concreto, como para qualquer outro operador é executado um conjunto de duas instruções SQL específicas: uma instrução que determina o número de itens do conjunto consequente ou antecedente (conforme o operador) e uma outra que identifica as regras de associação que obedecem à "transformação" requerida pelo utilizador. Existe uma terceira instrução SQL independente do operador, que permite obter as características das regras identificadas pela segunda instrução SQL (itens, conjunto antecedente, conjunto consequente, valor de suporte e valor de confiança). Esta instrução permitirá construir a página (conjunto de regras) primeiro sob a forma de tabela e, posteriormente, sob a forma gráfica.

Capítulo IV Avaliação do método proposto

Este capítulo pretende avaliar o sistema PEAR em termos de ferramenta de suporte à metodologia de operadores apresentada nesta dissertação. Para aferir da utilidade e capacidade de funcionamento do PEAR, foram utilizados dois conjuntos de regras, obtidos a partir dos registos de acessos ao *site* do Instituto Nacional de Estatística (http://www.ine.pt). Este tipo de dados são vulgarmente conhecidos por *site access Logs*[16] e contêm informação sobre os acessos efectuados por todos os utilizadores do *site*, num determinado período de tempo. Dos conjuntos utilizados no PEAR foram eliminadas quaisquer referências a entidades colectivas ou individuais. Desta forma, foi possível utilizar uma grande quantidade de dados reais, salvaguardando o necessário anonimato dos utilizadores deste *site*.

Refira-se, ainda, que a análise das consultas de utilizadores a páginas de *sites* está a assumir, actualmente, uma grande importância junto de grandes empresas que possuem páginas na Internet (VOGELMAN, 2001). Partindo da forma como um utilizador navega num *site* e o tipo de páginas que consulta, é possível extrair informação útil sobre, por exemplo, as suas preferências. Este tipo de análise é utilizado por empresas como a conhecida livraria norte-americana *Amazon* (http://www.amazon.com) ou pela cadeia portuguesa de hipermercados *Continente* (http://www.continente.pt). Ambas as empresas fornecem aos utilizadores um conjunto de páginas adaptadas às suas preferências, sem que o utilizador se aperceba, procurando assim tornar mais apelativa a sua navegação no *site*. No caso da *Amazon*, por exemplo, sempre que determinado utilizador ou cliente procura informação ou adquire um determinado livro, imediatamente recebe mais informação acerca de livros semelhantes que lhe possam interessar. Deste modo, procura-se influenciar no momento crítico de uma compra de um produto, a decisão do cliente.

[16] *Site Logs* : ficheiros criados pelo serviço de Internet de um servidor, que contém informação detalhada sobre a navegação dos utilizadores pelo respectivo *website*. Pode conter, não só o conjunto de páginas a que determinado utilizador acedeu, como também que informação enviou, que imagens observou, etc.

A adaptação do conteúdo das páginas de um *site*, consultadas por determinado utilizador, é baseada em informações obtidas a partir dos *Logs* de utilizadores anteriores que possuam características semelhantes ao referido utilizador. Este tipo de adaptação de conteúdos às características/preferências de um utilizador designa-se normalmente por *personalização*.

1 Exemplo de utilização do PEAR

Para a demonstração da navegação pelo espaço de regras escolheu-se um conjunto de 211 regras produzidas a partir das consultas efectuadas ao *site* do Instituto Nacional de Estatística. As regras associam as consultas efectuadas aos temas estatísticos, por utilizador e por sessão, ao *site* do Instituto Nacional de Estatística, entre Junho de 1997 e Dezembro de 1999. Estas regras de associação foram produzidas pelo algoritmo *Apriori* e exportadas para um ficheiro em formato texto. Procedeu-se então a uma importação para Microsoft Excel, passo intermédio necessário para a conversão final, das regras, para um documento PMML. O documento PMML que contém o modelo de regras de associação pode ser consultado no anexo deste texto. Após a criação do documento PMML, as regras de associação estão prontas a serem utilizadas pelo sistema PEAR.

O interface deste sistema é bastante simples e intuitivo. As quatro opções disponíveis, na parte superior do ecrã, quando a aplicação é invocada no *web-browser* são:

- *Home* – a página ou ecrã de entrada no PEAR;
- *Input* – a página que permite definir qual o documento PMML a aceder;
- *Visualize* – a página que permite efectuar a navegação pelo espaço de regras obtidas na página *Input;*
- *Export* – esta página irá permitir exportar um determinado agrupamento de regras, definido na página *Visualize*, para o formato PMML.

1.1 Definição do documento PMML a utilizar

O passo que antecede a navegação, propriamente dita, pelo espaço de regras consiste na identificação do documento que contém as regras de associação a analisar. O utilizador deve introduzir o endereço (endereço local ou endereço Internet) completo para o documento PMML. Um endereço Internet perfeitamente válido é, por exemplo, a expressão “http://www.3pontos.com/pear/imp/infoline.xml”. Se, em vez de se tratar de

um documento localizado num servidor da Internet, se tratasse de um documento localizado no computador do utilizador, o endereço a introduzir poderia ser, por exemplo, "http://localhost/pear/imp/infoline.xml". Na figura seguinte, após a introdução do endereço do documento PMML e do utilizador ter pressionado o botão "get it", aparece na página um conjunto de instruções em linguagem XML que exprimem o modelo de regras de associação obtido. Importa notar que o documento PMML deve ser um documento XML válido, caso contrário, o utilizador recebe uma mensagem de erro alertando-o para o facto de se tratar de um documento com erros de sintaxe.

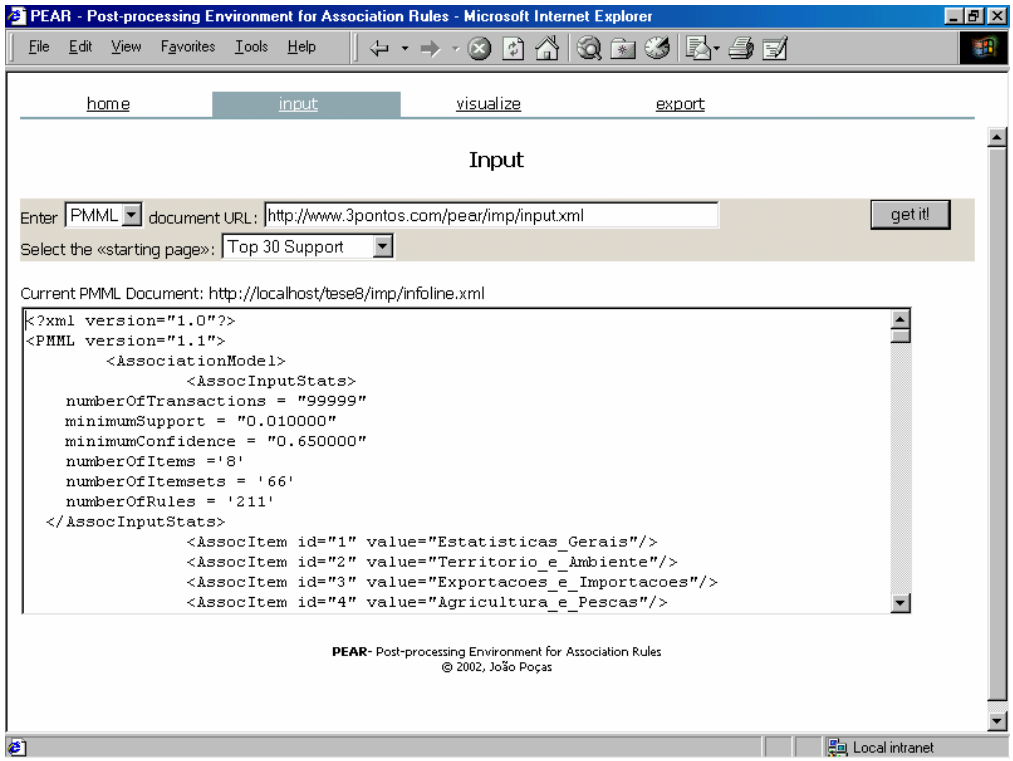


Figura 40 Leitura de um modelo de regras de associação em PMML, no PEAR.

É ainda nesta página que o utilizador deve seleccionar o tipo de primeira página que deseja visualizar, ou seja, a página com que irá iniciar a sua navegação. Esta opção permite definir três tipos de primeira página:

- *Top 30 Support*: esta primeira página será constituída por, no máximo, 30 regras de associação, ordenadas pelo valor do seu suporte (das regras com maior suporte para as regras com menor suporte);
- *Top 30 Confidence*: define uma primeira página constituída por, no máximo, 30 regras de associação, ordenadas pelo valor da sua confiança (das regras com maior confiança para as regras com menor confiança);
- *All rules*: a primeira página será constituída por todas as regras de associação que fazem parte do conjunto de regras inicial. Para grandes conjuntos de regras, esta não será uma boa página inicial.

A escolha de uma página inicial adequada ao problema em análise pode ser decisiva para a melhor compreensão do mesmo. Alguns dos estudos abordados no primeiro capítulo deste texto poderiam ser aplicados no PEAR, permitindo uma variedade de "primeiras páginas". A utilização de conceitos como a surpresa de uma regra ou regras de senso comum (HUSSAIN et al., 1999) permitiriam, por exemplo, definir uma primeira página com o conjunto de regras mais inesperadas. Quaisquer medidas que permitam caracterizar objectivamente regras de associação, podem ser utilizadas na definição de uma página inicial. Após a selecção da página inicial, o utilizador está apto a visualizar a "sua primeira página" de regras de associação e dar início, desse modo, à navegação pelo espaço de regras, de página em página, utilizando o método dos operadores.

1.2 Página inicial

A visualização do conjunto de regras da página inicial (página de partida para a navegação) é disponibilizada através da opção *visualize*, existente também na parte superior do ecrã do PEAR. No PEAR, uma página pode ser interpretada como um conjunto de regras proveniente do espaço de regras de associação. O analista utiliza o método dos operadores de regras de associação para navegar ou consultar diferentes páginas, que significa o mesmo que consultar diferentes conjuntos de regras de associação. Por esta razão, facilmente se pode entender a navegação no espaço de regras

de associação através dos operadores, como uma navegação através de páginas interligadas na Internet.

Na figura seguinte pode observar-se uma parte do conjunto de regras que constituem a página inicial obtida do modelo de 211 regras de associação provenientes do *site* do Instituto Nacional de Estatística. Para o caso em estudo, escolheu-se o tipo de página "*Top 30 Support*" como página inicial.

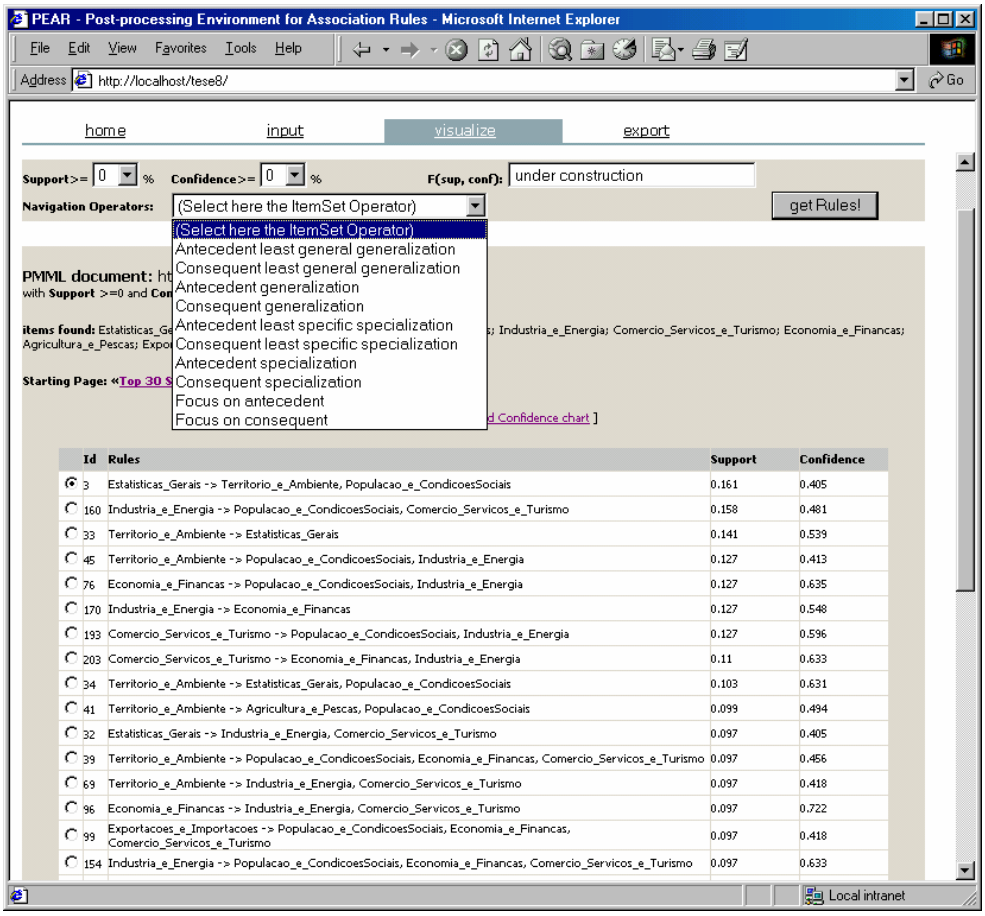


Figura 41 Conjunto de regras iniciais ou página inicial do PEAR.

Na página inicial, o utilizador é colocado perante diversas alternativas de navegação, apesar do principal realce incidir, obviamente, sobre os diferentes operadores de regras (*navigation operators*). Na figura anterior (figura 41), observam-se, claramente, os diversos operadores de regras disponíveis. Estes operadores podem ser aplicados a

qualquer regra que o utilizador entenda conveniente para a sua análise. Assim, antes de aplicar qualquer um dos operadores, é necessário que o utilizador seleccione a regra de associação que lhe interessa "transformar". Para além de uma navegação específica, definida por intermédio dos operadores e orientada para uma regra seleccionada pelo utilizador, é possível ainda filtrar o conjunto de regras disponíveis na página, visualizando apenas as regras que possuem um determinado valor de suporte e de confiança. Estes filtros estão disponíveis no topo da página, sob a forma de "caixas de selecção", permitindo definir valores de confiança e suporte entre 0 e 100 por cento.

Uma outra funcionalidade, colocada à disposição do analista, que pode facilitar a análise ou a selecção do caminho a seguir, é a componente visual. Esta componente pode ser acedida através de duas opções (*Rules chart* e *Support and Confidence chart*), existentes em cada uma das páginas de visualização de regras de associação. Estas duas formas de visualização, apesar da sua simplicidade, permitem observar, de um modo mais rápido, a distribuição, em termos de valores de confiança e suporte, do conjunto de regras disponíveis na página em análise.

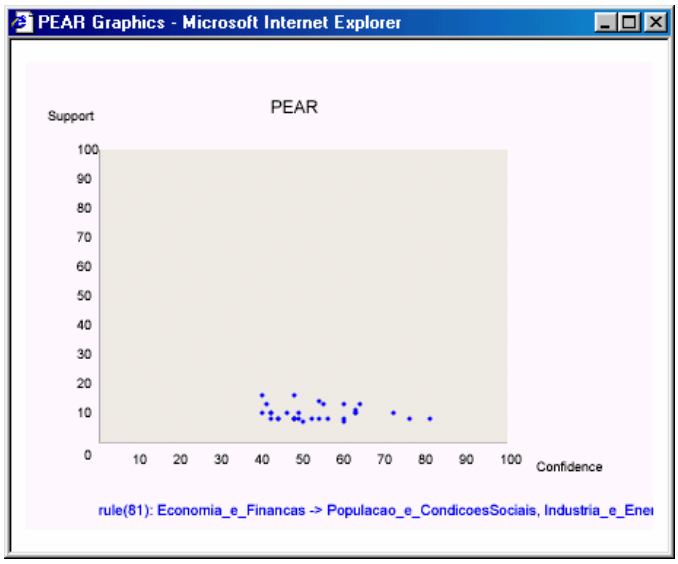


Figura 42 Gráfico *Rule Chart* produzido pelo conjunto de regras da página inicial.

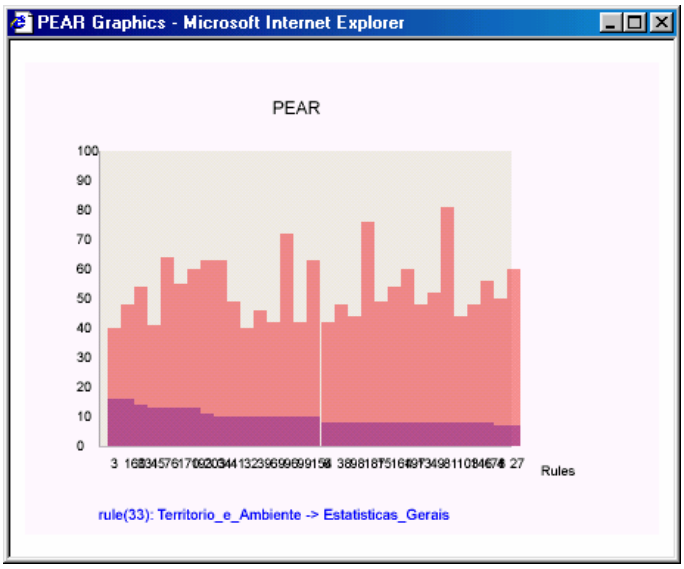


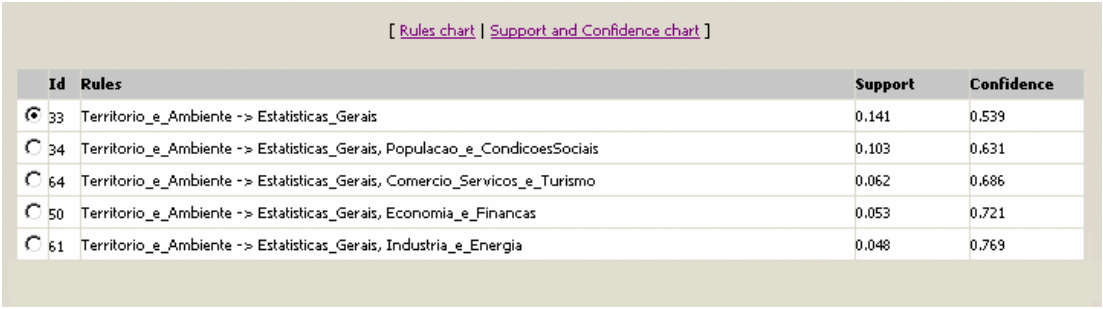
Figura 43 Gráfico *Confidence and Support* Chart produzido pelo conjunto de regras da página inicial.

Pela análise das figuras anteriores, que correspondem ao conjunto de regras disponíveis na página inicial considerada anteriormente, é visível a existência de valores de suporte bastante baixos (entre os 10% e os 20%) sendo que a maior parte dos valores de confiança se concentram entre os 40% e os 60%. A componente visual, ou gráfica, permite ainda utilizar os gráficos para identificar a regra que corresponde a determinado ponto (figura 42) ou a regra que corresponde a determinada barra (figura 43).

1.3 Navegando pelo espaço de regras

Partindo do conjunto de regras obtidas na página inicial, o analista pode percorrer o espaço de regras de associação, aplicando operadores (e filtros) ao referido conjunto. Assumindo que o analista considera as regras que possuem o tema estatístico "Territorio_e_Ambiente" interessantes de analisar, e procuram, mais especificamente, relacionar este tema estatístico com o tema "Estatisticas_Gerais", seleccionou-se a terceira regra da página inicial, com o número de identificação 33, aplicando-se o operador *ConsS* (*Consequent Specialization*). Com este operador, obtêm-se regras com consequentes que possuem um ou mais itens para além dos que constituem o consequente da regra seleccionada. com um suporte inferior, mas uma confiança mais

elevada que a regra corrente. Procura-se, assim, obter um conjunto de regras mais específico que a regra referida.

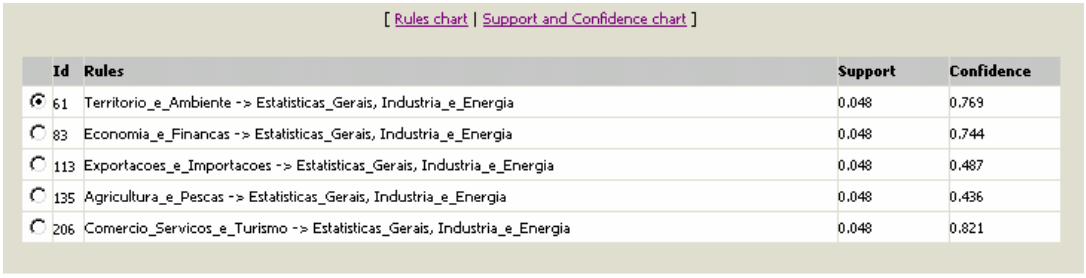


[Rules chart | Support and Confidence chart]

	Id	Rules	Support	Confidence
◉	33	Territorio_e_Ambiente -> Estatisticas_Gerais	0.141	0.539
○	34	Territorio_e_Ambiente -> Estatisticas_Gerais, Populacao_e_CondicoesSociais	0.103	0.631
○	64	Territorio_e_Ambiente -> Estatisticas_Gerais, Comercio_Servicos_e_Turismo	0.062	0.686
○	50	Territorio_e_Ambiente -> Estatisticas_Gerais, Economia_e_Financas	0.053	0.721
○	61	Territorio_e_Ambiente -> Estatisticas_Gerais, Industria_e_Energia	0.048	0.769

Figura 44 Conjunto de regras (ou página) resultante da aplicação do operador ConsS.

Como se pode observar pela figura anterior, a página resultante da aplicação do operador *ConsS*, apresenta um conjunto de apenas cinco regras, tornando a sua interpretação mais fácil e rápida que a página inicial. Observando os valores de suporte e confiança apresentados para cada regra, verifica-se, como seria de esperar, que as regras obtidas apresentam valores mais altos de confiança, em detrimento de suportes mais baixos. Das regras obtidas, verifica-se, por exemplo, que o tema estatístico "Industria_e_Energia" não está significativamente associado com "Territorio_e_Ambiente" nem "Estatisticas_Gerais", pelo que se poderia optar por aplicar um novo operador à regra "Territorio_e_Ambiente -> Estatisticas_Gerais, Industria_e_Energia". Supondo que o analista aplica o operador *FCons* (*Focus on Consequent*), as regras a obter poderão ter quaisquer antecedentes, mantendo o mesmo consequente.



[Rules chart | Support and Confidence chart]

	Id	Rules	Support	Confidence
◉	61	Territorio_e_Ambiente -> Estatisticas_Gerais, Industria_e_Energia	0.048	0.769
○	83	Economia_e_Financas -> Estatisticas_Gerais, Industria_e_Energia	0.048	0.744
○	113	Exportacoes_e_Importacoes -> Estatisticas_Gerais, Industria_e_Energia	0.048	0.487
○	135	Agricultura_e_Pescas -> Estatisticas_Gerais, Industria_e_Energia	0.048	0.436
○	206	Comercio_Servicos_e_Turismo -> Estatisticas_Gerais, Industria_e_Energia	0.048	0.821

Figura 45 Conjunto de regras (ou página) resultante da aplicação do operador FCons.

Na figura anterior (figura 45) observam-se as regras que possuem as associações mais fortes com os temas “Estatisticas_Gerais” e “Industria_e_Energia”. O processo de navegação continuará a ser iterativo, permitindo ao analista percorrer um determinado caminho, pelo espaço de regras de associação, que considere interessante. Desta forma iterativa, o analista irá obtendo um conhecimento aprofundado sobre o conjunto de regras (e sobre os dados em análise).

2 Análise de desempenho

Quando se procurou implementar um sistema para funcionar via Internet, de imediato se levantaram algumas dúvidas quando à capacidade de funcionamento e celeridade de tal sistema. As dúvidas aumentaram quando se pretendeu criar um sistema que fosse independente quer do *web-browser* quer do sistema operativo utilizado. O PEAR possui as três características em simultâneo: funciona através da Internet, funciona nos *web-browsers* utilizados por mais de 90% dos utilizadores de Internet e é independente da máquina que o utiliza. Sendo assim, não é fácil assegurar uma total capacidade de funcionamento e celeridade de um sistema deste género. Certamente que um sistema como o PEAR, implementado numa plataforma não Internet, por exemplo em ambiente *Microsoft Windows* ou ambiente *Linux*, forneceria, à partida outras garantias quanto à velocidade de processamento e de manipulação de um enorme número de regras. No entanto, perder-se-iam as potencialidades próprias que caracterizam um sistema que funciona num ambiente Internet: portabilidade, independência da máquina e utilização remota. Desta forma, importa então aferir da capacidade de resposta em termos de velocidade de processamento de um sistema desta natureza, quando confrontado com um elevado número de regras de associação. Para os testes realizados com o PEAR foi utilizado um computador com um processador *Intel Celeron* (*pentium II*) a 266Mhz com 128Mb de memória Ram.

1.1 Um caso de milhares de regras

O conjunto de regras de associação utilizadas para este caso referem-se às consultas efectuadas entre Junho de 1997 e Dezembro de 1999 ao *site* do Instituto Nacional de Estatística. Sendo que, mais do que uma análise às regras de associação obtidas, pretendia-se conhecer a resposta do sistema a um elevado número de regras, produziu-se um conjunto de milhares de regras que não possuíam qualquer utilidade analítica. O

conjunto de regras, obtido através do algoritmo *Apriori*, foi gerado em formato CSV[17], sendo depois convertido para formato PMML, através de uma simples aplicação, desenvolvida especificamente para o efeito, em *Microsoft Visual Basic*. O documento PMML apresentava uma dimensão física de 815Kb; continha 124 itens diferentes (quadros estatísticos); possuía 1541 conjuntos diferentes de itens e armazenava um total de 6702 regras.

Este conjunto de regras foi utilizado em ambiente Internet e em ambiente local, por forma a possibilitar uma melhor avaliação do comportamento do sistema em diferentes ambientes. A resposta do sistema a estes dois tipos de ambientes está sintetizada no quadro seguinte.

Tabela 5 Tempos de resposta do sistema PEAR, para 6 702 regras de associação.

	Tempo de resposta (em segundos)		Relação (2)/(1)
	Ambiente Local (1)	Ambiente Internet (2)	
1. Carregamento e validação do documento PMML (efectuado sempre no PC do utilizador)	5s	42s	8,4
2. Inserção na Base de Dados (servidor)	840s	2 760s	3,2
3. Navegação (servidor e PC do utilizador)	2s	4s	2

Uma ilação que rapidamente se extrai, numa primeira observação do quadro anterior, é a de que o desempenho do sistema piora abruptamente quando o mesmo é utilizado através da Internet.

O tempo de resposta no primeiro ponto da tabela anterior - carregamento e validação do documento PMML - representa a soma do tempo de transferência do documento PMML do computador servidor para o computador cliente com o tempo de execução no *web-browser* (computador local). Sendo o processamento executado no computador cliente, quer esteja o utilizador a trabalhar em ambiente Internet ou em ambiente local, a diferença de valores obtida, neste caso, tem que ver sobretudo com a parcela

[17] CSV: significa *Comma Separated Values* em inglês, ou "valores separados por vírgula". É um formato de documento que pode ser facilmente criado e manipulado com uma aplicação do tipo da *Microsoft*

correspondente à transferência do documento entre computadores. Considerando que, existiu uma transferência de um documento cuja dimensão física ascendia aos 800 *kbytes*, facilmente se depreende que 42 segundos é um tempo bastante bom, apesar de corresponder a um tempo de processamento cerca de 8 vezes maior que o tempo de resposta local. No ambiente Internet, este tempo poderá ser diminuído se o utilizador utilizar uma ligação à Internet mais veloz, uma vez que o processo de ligação utilizado neste teste foi um simples *modem* de 56 *kbps*[18]. Numa experiência realizada em condições de ligação superiores (semelhantes a uma ligação ADSL[19]) o desempenho no ambiente Internet, quando se trata de carregamento em memória do documento PMML, assemelham-se ao carregamento local (cerca de 5 segundos).

No segundo ponto da tabela anterior, a dificuldade do sistema PEAR é também maior quando em ambiente Internet, se comparado com um funcionamento local: o processo de inserção das regras de associação na base de dados demora 2 760 segundos, cerca de 3,2 vezes mais que a funcionar localmente.

1.2 Um caso de centenas de regras

O conjunto de regras utilizado no ponto anterior deste capítulo serviu apenas como teste de robustez e avaliação de desempenho do sistema PEAR. Não apresentava informação útil para se proceder a uma análise real, uma vez que associava determinados itens (quadros estatísticos) sem considerar a estrutura em que se encontram esses quadros (publicações). Para efectuar uma navegação sobre um conjunto de regras, fazia todo o sentido que o mesmo contivesse informação aplicável a um caso real. Deste modo, para que o processo de navegação pelo espaço de regras fosse o mais real possível, recorreu-se a um conjunto de regras mais pequeno mas sem dúvida mais interessante e útil.

Excel.

[18] *Modem* 56 *kbps*: dispositivo electrónico que permite efectuar a ligação de um computador com a rede Internet; 56*kbps* representa uma velocidade de transmissão de dados de 56 *kilobits* por segundo.

[19] ADSL – significa Asymmetric Digital Subscriber Line, em inglês, e define uma tecnologia de comunicação de banda larga que usa linhas telefónicas comuns.

As consultas de quadros estatísticos, efectuadas por utilizadores de Internet, ao *site* do Instituto Nacional de Estatística entre Junho de 1997 e Dezembro de 1999, foram agrupadas por tema estatístico. Pode utilizar-se um certo paralelismo entre as compras de produtos de supermercado efectuados por um cliente, numa determinada data, e as consultas de quadros estatísticos efectuadas por um utilizador num determinado período de tempo. As compras efectuadas numa determinada data constituem uma transacção. As consultas efectuadas por um utilizador num certo período de tempo constituem uma sessão. As sessões são facilmente determinadas no ficheiro de *Logs* produzido pelo servidor do *site*. Assim, foi produzido um conjunto de 211 regras de associação que exprimem as associações entre os temas consultados. Interessa, à entidade em causa, conhecer os hábitos de consulta dos utilizadores do seu *site*. O principal objectivo desta análise é o de aproximar cada vez mais a informação estatística produzida das necessidades dos utilizadores e produtores das estatísticas.

Os tempos de resposta do sistema PEAR para este conjunto de regras foram também registados, para se averiguar a diferença de desempenho do PEAR perante um menor conjunto de regras. No quadro seguinte, pode-se observar os tempos obtidos.

Tabela 6 Tempos de resposta do sistema PEAR, para 211 regras de associação.

	Tempo de resposta (em segundos)		Relação (2)/(1)
	Ambiente Local (1)	Ambiente Internet (2)	
1. Carregamento e validação do documento PMML (efectuado sempre no PC do utilizador)	1s	6s	6
2. Inserção na Base de Dados (servidor)	13s	18s	1,3
3. Navegação (servidor e PC do utilizador)	2s	4s	2

Uma análise comparativa dos tempos de resposta do sistema, perante um conjunto de regras de menor dimensão (de 6 702 regras de associação no primeiro conjunto para 211 no segundo), aponta para uma diminuição significativa no tempo de inserção das regras na base de dados, quer em ambiente local, quer em ambiente Internet. Quando comparamos a velocidade de processamento do PEAR, para o mesmo conjunto de regras, em ambientes diferentes, os resultados são evidentes: a velocidade de

processamento do PEAR reduz-se significativamente quando este é utilizado via Internet. Por outro lado, verifica-se que o processo que ocupa um maior período de tempo tem que ver, uma vez mais, com a inserção de dados na base de dados.

A maior diferença (em termos percentuais) nos tempos de resposta do sistema, entre os dois ambientes, verifica-se no processo de carregamento para memória do documento PMML, à semelhança do que aconteceu na experiência com um conjunto de milhares de regras de associação. Também neste caso, a velocidade de ligação ou comunicação entre o computador do utilizador e o servidor de Internet tem um peso muito grande, no tempo de resposta do PEAR. Apesar do documento PMML possuir uma dimensão física de apenas 27 *Kbytes*, a transmissão dos dados e as comunicações de confirmação que os protocolos de comunicação exigem, obrigam a um tempo de resposta mais dilatado.

Importa ainda referir que, na fase de navegação, os tempos de resposta do sistema se manteve o mesmo, para os dois conjuntos de regras. Estes resultados são idênticos, uma vez que o tempo de resposta do sistema, na fase de navegação entre conjuntos de regras, depende da quantidade de informação visualizada (transmitida) em cada página, ou seja, do número de regras de associação contidas em cada página.

3 Aspectos positivos

Após os resultados quanto ao desempenho obtido no ponto anterior e o exemplo de utilização realizado, faz todo o sentido apontar os pontos fortes da metodologia apresentada nesta dissertação. Em termos globais, a metodologia baseada em operadores, e mais concretamente o PEAR, cumpre o seu objectivo, ou seja, assume-se como um contributo para a interpretação de um grande número de regras de associação.

Através do exemplo apresentado no início deste capítulo, pode constatar-se da facilidade de compreensão e utilização dos operadores, que permitem uma navegação pelo espaço de regras visualizando um conjunto de regras de cada vez. Desta forma, o utilizador não é confrontado com um grande número de regras para analisar que o podem deixar desorientado, permitindo, pelo contrário, uma análise mais direccionada.

O interface simples do PEAR permite uma aprendizagem rápida, possibilitando assim, uma utilização imediata por parte dos utilizadores. À simplicidade do interface está associado o ambiente em que o sistema foi desenvolvido: a Internet. Deste modo, ficam asseguradas algumas vantagens relacionadas com a não necessidade de instalação deste sistema, uma vez que pode ser utilizado remotamente, através de um simples *web-browser*. Por outro lado, facilita a partilha de informação e opiniões entre diferentes utilizadores fisicamente distantes.

O PEAR, para além de implementar a metodologia dos operadores sobre regras de associação, implementa também uma componente gráfica que constitui um complemento à análise normal em forma de tabela. Esta componente não está tão desenvolvida como alguns estudos desenvolvidos por outros investigadores (WONG, P. C. et al., 1999), mas poderá, no futuro, ser melhorada. Uma outra facilidade disponível no PEAR tem que ver com o formato dos modelos de regras de associação utilizados. Estes modelos devem estar representados em formato PMML, um formato *standard* para representação de modelos de *data mining*.

4 Aspectos negativos

No decorrer das experiências realizadas, assim como durante o processo de desenvolvimento do PEAR, foram perceptíveis alguns problemas e alguns pontos fracos do sistema. Apesar de se ter procurado eliminar os problemas encontrados à medida que se ia desenvolvendo o sistema, outros, contudo, persistiram até hoje. Para além disso, algumas questões tais como a implementação de operadores que se aplicam a conjuntos de regras de associação ficaram ainda por implementar. Neste capítulo, serão abordados e analisados, com maior pormenor, alguns problemas assim como algumas questões que poderiam completar ou complementar o sistema PEAR. Numa perspectiva de continuidade de desenvolvimento deste sistema, apontam-se ainda algumas medidas que poderiam optimizar e até solucionar algumas carências evidenciadas pelo PEAR.

4.1 Limitações

As experiências realizadas com o PEAR parecem corroborar a ideia, lançada no início deste capítulo, de que a implementação de um sistema via Internet pode ter algumas limitações de funcionamento. Sem dúvida que o principal problema que pode ser apontado ao PEAR é o seu desempenho na manipulação de mais do que uma dezena de milhares de regras, uma vez que, para centenas de regras, as diferenças de tempos de resposta quando em ambiente local e quando em ambiente Internet não são significativas. As exigências dos protocolos de comunicação entre computadores (como é o caso do protocolo http e do protocolo ftp[20]), quanto às frequentes verificações de transferências de informação em perfeitas condições, acabam por atrasar a execução das tarefas do sistema.

O processo mais nefasto para a perda de desempenho em ambiente Internet é sem dúvida a execução de inserções na base de dados (inserção de regras de associação, de

[20] Ftp: significa *File Transfer Protocol*, em inglês; recurso que possibilita a transferência de ficheiros de um servidor ou computador da Internet para o computador do utilizador e vice-versa.

itens e de conjuntos de itens), após uma interpretação do documento PMML inicial. Tratando-se de instruções, na sua maior parte, de instruções de SQL, torna-se evidente que um SGBD mais eficiente do que o *Microsoft Access* ajudará. Se, esta substituição por um sistema de gestão de base de dados mais robusto, fosse acompanhada pela utilização de uma ligação à Internet mais veloz, ficaria assegurado um tempo de execução bastante menor. O processamento de leitura e validação do documento PMML, assim como o processo de manipulação dos conjuntos de regras de associação, recorrendo aos operadores de regras, beneficiaria de igual modo com a melhoria de ligação e melhoria de conexão com a base de dados.

4.2 Principais problemas

O principal desafio, lançado na génese deste trabalho, era o de implementar um sistema que pudesse demonstrar as capacidades de um novo método, para navegação sobre um grande conjunto de regras de associação. O PEAR parece ter superado este desafio: não só aplica, na prática, o método dos operadores como forma de navegação sobre um conjunto de regras, como vai um pouco mais longe, permitindo, por exemplo, utilizar a visualização gráfica como complemento à utilização dos operadores. Importa realçar que, no decorrer do desenvolvimento do sistema, algumas características suplementares foram sendo adicionadas, mas outras, por inerência, ficaram ainda incompletas ou por implementar. Nos pontos seguintes, estão enumeradas algumas questões que evidenciam alguma fragilidade ou ficaram por implementar no PEAR e que devem ser alvo de desenvolvimentos futuros.

Página inicial

O conjunto inicial de regras, a partir da qual se efectua a navegação, assume um papel crucial para uma análise mais profunda. No PEAR apenas foram implementados três tipos simples de página inicial: um conjunto máximo de 30 regras ordenadas pelo valor de suporte, um conjunto máximo de 30 regras ordenadas pelo valor de confiança e o conjunto de todas as regras. Seria interessante considerar outros tipos de página inicial. Alguns estudos anteriores, que exploravam diferentes métodos de medir o interesse das

regras de associação, podem ser aplicados a esta página inicial. Ao utilizador deixar-se-ia a liberdade de escolher a página que melhor servisse os seus propósitos.

Definição de "operadores à medida"

O interface já desenvolvido prevê a utilização de operadores definidos pelo utilizador, possibilitando uma navegação mais "à medida das preferências do utilizador". O utilizador pode ter interesse em visualizar as regras que possuam determinadas propriedades (suporte, confiança, *lift*, *laplace*, etc.). No entanto, esta funcionalidade ainda não está implementada no sistema actual.

Visualização gráfica

Apesar dos esforços de concretização de uma forma de consulta gráfica de regras de associação, a solução desenvolvida não está perfeita, evidenciando algumas carências, ao nível, por exemplo, da interactividade, pelo que existe também algum trabalho a desenvolver nesta área. No entanto, a tecnologia utilizada - *Scalable Vector Graphics* - parece fornecer boas indicações para futuros desenvolvimentos: possibilidade de aliar a qualidade de gráficos vectoriais ao recurso à interactividade com o utilizador.

Filtros de regras

Na navegação por um determinado espaço de regras, o PEAR possibilita uma "filtragem" das regras visualizadas em qualquer página dessa navegação, recorrendo a restrições de valores de confiança e suporte. Seria interessante utilizar outro tipo de filtros, quer considerando medidas já conhecidas, como é o caso do *lift*, *gain, laplace* e outras, quer considerando funções definidas pelo próprio analista.

Representação de diversas medidas

O método de navegação do PEAR, baseado na visualização de páginas, que representam conjuntos de regras, permite observar, para cada regra de associação, os valores correspondentes de suporte e confiança. Seria também interessante, como complemento, poder visualizar, para cada regra, os valores de outras medidas alternativas.

Múltiplos utilizadores

O ambiente Internet caracteriza-se por facilitar a existência de ambientes multi-utilizador, ou seja, permitir que um mesmo sistema seja utilizado por diferentes utilizadores, simultaneamente. O ambiente PEAR, apesar de permitir a utilização de diversos utilizadores simultâneos, restringe esta utilização ao mesmo modelo de regras de associação, impedindo a possibilidade de se analisar, em simultâneo, dois problemas paralelos. No entanto trata-se de uma limitação facilmente ultrapassável, sendo necessário, para que tal seja possível, adaptar as tabelas do modelo de dados para que permitam manipular, em simultâneo, vários modelos diferentes.

Capítulo V Conclusões

A descoberta de regras de associação, introduzida por Agrawal (AGRAWAL et al., 1993), é uma técnica de *data mining* que permite extrair conhecimento a partir de grandes volumes de dados. A interpretação e análise de regras de associação, geradas a partir de problemas que envolvem pequenos volumes de dados, não apresenta, geralmente, qualquer dificuldade, uma vez que o espaço de regras de associação é normalmente pequeno. O mesmo não se passa quando o volume de dados é de tal ordem elevado, que o número de regras de associação geradas facilmente ultrapassa o humanamente compreensível. Este motivo levou à definição de alguns métodos que o procuram contrariar.

Esta dissertação debruçou-se sobre o estudo de uma nova metodologia que pretende facilitar a interpretação e análise de grandes volumes de regras de associação: o método dos operadores de regras de associação. Neste capítulo será feita uma retrospectiva aos trabalhos desenvolvidos nesta área, com particular relevo para o método dos operadores, assim como abordará perspectivas para futuros desenvolvimentos.

1 Retrospectiva do trabalho realizado

Apesar de ser relativamente fácil compreender a informação que as regras de associação transmitem, quando o seu número ascende a alguns milhares a interpretação das mesmas torna-se bastante difícil. Este motivo levou a que se desenvolvessem diversos estudos que procuravam formas de ajudar na interpretação de regras de associação. Alguns investigadores preocuparam-se com a introdução de melhorias nos algoritmos de geração de regras, recorrendo a métodos de resumo e agrupamento (TOIVONEN et al., 1995 e LIU et al., 1999b). Estes métodos, apesar de reduzirem o número de regras geradas, não são, por si só, suficientemente eficazes para garantir uma redução de regras que permitisse uma eficaz interpretação das regras.

Elaboraram-se outros estudos na área da visualização gráfica de regras de associação, como forma de apoiar a interpretação das mesmas. A componente gráfica, permite analisar mais rapidamente um maior número de regras de associação (KLEMETTINEN, M. et al*,* 1994 e LIU et al., 1999a). Porém, quando o número de regras a analisar atingem um valor bastante elevado, torna-se quase infrutífero representá-las graficamente (WONG, P. C. et al., 1999).

Uma outra corrente de investigadores procurou focar o seu estudo nas propriedades das regras de associação que permitiam medir o seu interesse. Os diversos estudos nesta área, concluíram que, apesar de se poderem encontrar medidas objectivas que caracterizem uma determinada regra, o interesse da mesma, em última análise, está inerente à subjectividade de quem a estuda. Foram também criados alguns conceitos que ajudam a identificar as regras mais interessantes, tais como a existência de regras inesperadas (LIU et al., 1999a), regras de "senso comum" (SAHAR, S., 1999), "aplicabilidade" (HUSSAIN et al., 1999), entre outro tipo de regras.

A utilização de sistemas de gestão de bases de dados, quer na descoberta, quer na ajuda à interpretação de regras de associação, foi também um caminho procurado por diferentes investigadores. Foram criadas linguagens inspiradas na linguagem SQL, por

forma a facilitar a selecção de regras de associação. Estas linguagens permitiam produzir, rapidamente, novos conjuntos de regras de associação que satisfizessem determinadas condições (MORZY, T. e M. ZAKRZEWICZ, 1997). A grande vantagem do armazenamento de modelos de regras de associação em sistemas de bases de dados, parece reunir consenso entre os diferentes investigadores: permite uma análise mais rápida e adaptável ao tipo de cenário ou analista.

Pela revisão dos diferentes estudos realizados, facilmente se depreende que a interpretação de um número elevado de regras de associação não tem uma resolução única. No entanto, qualquer dos investigadores abordados no texto conseguiu contributos assinaláveis para minorar os efeitos do problema.

O método dos operadores de regras de associação, apresentado nesta dissertação, pretende ser também uma contribuição para a ajuda à interpretação de um grande número de regras. O sistema PEAR surgiu como uma forma de implementação deste método. O PEAR, além de implementar o método dos operadores incorpora algumas técnicas propostas por estudos referidos neste texto. É o caso da visualização gráfica de regras de associação. No caso do PEAR, a componente gráfica assume um papel de complemento à análise das regras de associação pela utilização de operadores. Uma ideia semelhante, aliás, havia já sido defendida por KLEMETTINEN (1994) que propunha uma ferramenta gráfica (*Rule Visualizer*) para complementar uma metodologia de filtros de regras.

O sistema PEAR ajuda o utilizador a navegar pelo espaço de regras de associação, recorrendo aos operadores, permitindo a visualização de um conjunto de regras de cada vez. Cada um destes conjuntos de regras é representado por uma página *web* dinâmica. O utilizador pode, desta forma, navegar de uma página para outra aplicando um operador à sua escolha. A escolha do ambiente Internet para desenvolvimento do PEAR deve-se também à popularidade que este meio de comunicação alcançou nos dias de hoje. Por um lado, permite diminuir o tempo de aprendizagem do utilizador que um sistema deste género exigiria, se desenvolvido noutro ambiente, uma vez que apresenta um interface simples de páginas *web*. Por outro lado, adquire as características próprias

de um ambiente desta natureza: portátil, independente da máquina em que está a ser utilizado e funcionamento remoto.

No PEAR acabaram por ser implementadas outras componentes que lhe atribuem funcionalidades que vão para além da simples navegação pelo espaço de regras. Deste modo, este sistema permite utilizar modelos de regras de associação provenientes de diferentes aplicações informáticas ou algoritmos, desde que estejam representados na linguagem PMML. Por outro lado, e para retirar proveito das capacidades de manipulação de dados dos sistemas de gestão de bases de dados, o modelo de regras de associação é, numa fase posterior, armazenado numa bases de dados. Ao contrário de outros estudos desenvolvidos nesta área, que recorriam a operadores e a linguagens novas (MEO, R. et al, 1996 ou HAN, J. et al., 1996), optou–se por recorrer à linguagem *standard* dos sistemas de gestão de bases de dados (a SQL) sem que o utilizador tenha de aprender uma nova linguagem, já que as instruções SQL são construídas automaticamente. As regras de associação que constituem as páginas do espaço de regras são obtidas através da construção automática de simples consultas em SQL à base de dados.

2 Limitações e Perspectivas

Um sistema como o PEAR, como qualquer outro sistema implementado em ambiente Internet, tem limitações associadas às características do próprio ambiente, de entre as quais se destaca a velocidade de comunicação entre computador local e o computador remoto (servidor que possui o sistema PEAR instalado). No entanto, considera-se que as vantagens que um sistema como este tem em estar acessível de qualquer parte do planeta, através da Internet, supera as desvantagens de poder ser, em algumas situações, um pouco lento no funcionamento. Apesar de se tratar de um sistema vocacionado para um funcionamento via Internet, o PEAR pode funcionar perfeitamente num computador local (com sistema operativo da família *Microsoft Windows*), possibilitando, desta forma, ganhos de velocidade assinaláveis.

Uma característica importante, se não mesmo decisiva, de um sistema que implementa o método dos operadores tem que ver com a página inicial. Esta página, que representa o conjunto de regras inicialmente visualizado pelo analista, deverá ser, na medida do possível definida pelo utilizador, uma vez que é a porta de entrada para as restantes páginas ou conjuntos de regras de associação do espaço total de regras. O PEAR apresenta apenas, três tipos simples de página inicial (*Top 30 Support*, *Top 30 Confidence* e *All rules*). Outros tipos de página inicial que considerassem, por exemplo, alguns estudos revistos neste texto, poderiam ser implementados de futuro, permitindo, assim, seleccionar a página inicial que melhor se adequasse ao problema em análise.

Outro aspecto importante, que poderá ser corrigido de futuro, prende-se com o facto do sistema PEAR não permitir um funcionamento multi-análise, ou seja, embora possibilite a análise simultânea do mesmo problema por diversos utilizadores, não permite que diversos utilizadores analisem diferentes problemas ao mesmo tempo. Esta limitação pode ser facilmente ultrapassada com uma simples adequação do modelo de base de dados utilizado, bastando para isso a atribuição de um identificador a cada modelo de regras utilizado, e correspondente alteração das rotinas que implementam os operadores (consultas à base de dados).

Apesar de não ser um requisito inicial deste trabalho, a possibilidade de exportação de regras (determinados conjuntos de regras) para o formato PMML teria todo interesse em ser implementada. Não o foi, embora o seu enquadramento esteja já definido no interface actual do PEAR, por forma a poder ser implementado no futuro.

A maior parte das empresas que desenvolvem as aplicações de extracção de conhecimento, que permitem gerar modelos de regras de associação, ainda não tomaram o formato PMML como uma necessidade urgente de satisfação dos seus clientes e utilizadores, embora muitas tenham aderido ao consórcio que define o *standard* PMML. Enquanto este formato não se generalizar às diversas aplicações informáticas, torna-se essencial garantir a conversão dos modelos de algumas dessas aplicações (normalmente exportáveis para ficheiros em formato texto) para o formato PMML. Neste sentido, foi criado um pequeno programa protótipo que permite converter um modelo de regras de associação, expresso em formato texto, para o correspondente modelo em formato PMML. No entanto, faria todo o sentido integrar este protótipo no ambiente PEAR.

3 Considerações Finais

A descoberta de regras de associação é uma técnica de extracção de conhecimento com bastante aplicabilidade a questões concretas, mas que coloca algumas dificuldades ao nível da sua análise.

Esta dissertação propõe e implementa um método de navegação pelo espaço de regras, de um qualquer modelo de regras de associação. Os conjuntos de regras que constituem o espaço de regras de associação são visualizados um de cada vez, e a transição de conjunto em conjunto processa-se através da aplicação de operadores sobre as regras ou sobre o conjunto. Deste modo, procura-se facilitar a interpretação de modelos que possuam um grande número de regras de associação. A utilização de modelos de regras de associação em formato PMML, proporciona ao sistema implementado uma independência funcional face às actuais aplicações informáticas de extracção de regras. Foram ainda implementadas algumas componentes adicionais, baseadas em técnicas desenvolvidas em estudos anteriores, como é o caso da componente base de dados e da componente gráfica.

A avaliação ao desempenho do PEAR demonstrou a viabilidade da utilização desta ferramenta na análise de um conjunto elevado de regras de associação, quer em termos de tempo de resposta do sistema, quer em termos de facilidade de navegaçaõ pelo espaço de regras. Considera-se que, apesar do principal objectivo desta dissertação ter sido atingido, há ainda muito trabalho que pode (e deverá) ser desenvolvido no futuro, nomeadamente ao nível da expansão de algumas das funcionalidades do PEAR, no sentido de tornar este sistema um contributo claro para a interpretação de um elevado número de regras de associação.

Este trabalho é apoiado pela União Europeia (grant IST-1999-11.495 Sol-Eu-Net), fazendo parte do projecto POSI/2001/Class apoiado pela Fundação Ciência e Tecnologia na área da descoberta e pós-processamento de regras de associação (http://www.niaad.liacc.up.pt/~amjorge/projectos/class).

Referências Bibliográficas

Artigos

Agrawal, R., T. Imielinski e A. Swami (1993), "Mining association rules between sets of itens in large databases." in Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 207-216.

Agrawal, R. e R. Srikant (1994), "Fast algorithms for mining association rules in large databases." in Proceedings of the 20th International Conference on Very Large Databases, pp. 487-499.

Bayardo, R. e R. Agrawal (1999), "Mining the Most Interesting Rules", in Proceedings of the 5th ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.145-154.

Brin, S., R. Motwani e C. Silverstein (1997a), "Beyond Market Baskets: Generalizing Association Rules to Correlations", in Proceedings of the ACM SIGMOD International Conference on Management of Data.

Brin S., R. Motwani, J. Ullman, S. Tsur (1997b), "Dynamic Itemset Counting and Implication Rules for Market Basket Data", In Proc.of the 1997 ACM-SIGMOD Int'l Conf. on the Management of Data, 255-264.

Fayyad, U. M., G. Piatetsky-Shapiro e P. Smith. (1996), "Knowledge Discovery In Databases: An Overview. In Knowledge Discovery In Databases", eds. G. Piatetsky-Shapiro, and W. J. Frawley, AAAI Press/MIT Press, pp. 1-36, Cambridge, MA.

Goethals, B., J. V. den Bussche e K. Vanhoof (1998), "Decision support queries for the interpretation of data mining results", Extended version of a paper in Proceedings of the 2nd International Conference on Data Warehousing and Knowledge Discovery

Goethals, B. e J. V. den Bussche (1999), "A priori versus a posteriori filtering of association rules", in ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.

Han, J., Jiawei Han, Y. Fu, W. Wang, K. Koperski e O. Zaiane (1996), "DMQL: A Data Mining Query Language for Relational Databases", in Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data.

Hao, M. C., U. Dayal, M. Hsu, T. Sprenger e M. Gross (2000), "Visualization of Directed Association in e-Commerce Transaction Data", in HP Labs 2000 Technical Report.

Hipp, J., C. Mangold, U. Güntzer e G. Nakhaeizadeh (2002), "Efficient Rule Retrieval and Postponed Restrict Operations for Association Rule Mining", Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'02), Taipei, Taiwan, pp. 52-65.

Houtsma, M. e A. Swami (1993), "Set-oriented mining of association rules", Research Report RJ 9567, IBM Almaden Research Center, San Jose, California.

Hussain, F., H. Liu e H. Lu (1999), "On Relative Measure for Mining Interesting Rules", in Principles of Data Mining and Knowledge Discovery 4th European Conference.

Jorge, A., J. Poças e P. Azevedo (2002a), "Post-processing Environment for Browsing Large Sets of Association Rules", in ECML/PKDD-2002 presented in Workshop Visual Data Mining IDDM-2002, Finland.

Jorge, A., J. Poças e P. Azevedo (2002b), "Post-processing Operators for Browsing Large Sets of Association Rules", in Proceedings of Discovery Science 2002 Eds, Steffen lange, Ken Satoh, Carl H. Smith, Springer-Verlag, LNCS 534, 2002.

Klemettinen, M., H. Mannila, P. Ronkainen, H. Toivonen e A. Verkamo (1994), "Finding interesting rules from large sets of discovered association rules", in R. Nabil et al., editors, Proceedings of 3rd International Conference on Information and Knowledge Management, pp. 401-407.

Liu, B., W. Hsu, K. Wang, e S. Chen (1999a), "Visually Aided Exploration of Interesting Association Rules", Third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). Lectures Notes in Artificial Intelligence 1574, pp. 380-389

Liu, B., W. Hsu e Y. Ma (1999b), "Pruning and Summarizing the Discovered Associations", in Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data mining.

Ma, Y, B. Liu e C. K. Wong (2000), "Web for Data Mining: Organizing and Interpreting the Discovered Rules Using the Web", in SIGKDD Explorations, vol. 2, issue 1, pp.16-23.

Meo, R., G. Psaila e S. Ceri (1996), "A new SQL-like operator for mining association rules", in T.M. Vijayaraman et al., editors, Proceedings of the 22nd International Conference on very Large Data Bases, pp. 122-123.

Morzy, T. e M. Zakrzewicz (1997), "SQL-Like Language For Database Mining", 1st Int'l Conference on Advances in Databases and Information Systems, pp. 311-317, St. Petersburg.

Morzy, T, M. Wojciechowski e M. Zakrzewicz (2000), "Materialized Data Mining Views", in Principles of Data Mining and Knowledge Discovery, 4th European Conference, PKDD 2000, Lyon, France.

Padmanabhan, B. e A. Tuzhilin (1999), "Unexpectedness as a Measure of Interestingness in Knowledge Discovery", in Decision Supports Systems, pp.303-318.

Pasquier, N, Y. Bastide, R. Taouil e L. Lakhal (1998), “Pruning Closed Itemset Lattices for Association Rules”, in Actes des 14\iemes journées “Bases de données avancées”, pages 177-196.

Sahar, S. (1999), “Interestingness Via What Is not Interesting”, in Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 332-336.

Silberschatz, A. e A. Tuzhilin (1995), “On Subjective Measure of Interestingness in Knowledge Discovery”, in Proceedings of the First International Conference on Knowledge Discovery and Data Mining, pp. 275-281.

Silberschatz, A. e A. Tuzhilin (1996), “What maks patterns interesting in knowledge discovery systems"” in IEEE Transactions in Knowledge Discovery and Data Engineering, pp 970-974.

Srikant, R. e R. Agrawal (1995), “Mining Generalized Association Rules”, in Proceedings of the 21st VLDB Conference, pp. 407-419.

Srikant, R., Q. Vu e R. Agrawal (1997), “Mining association rules with item constraints”, in D. Heckerman et al., editors, Proceedings 3rd International Conference on Knowledge Discovery and Data Mining, pp. 66-73.

Toivonen, H., M. Klemettinen, P. Ronkainen, K. Hätönen e H. Mannila (1995), “Pruning and Grouping Discovered Association Rules”, in MLNet Workshop on Statistics, Machine Learning and Discovery in Databases, pp.47-52.

Vogelman, J. (2001), "Determining Web Usability Through an Analysis of Server Logs", A Thesis Presented to The Faculty of the School of Engineering and Applied Science - University of Virginia, EUA

Webb, G. (1995), “OPUS: an efficient admissible algorithm for unordered search”, in Journal of Artificial Intelligence Research, 3:431-465.

Webb, G. (2000), “Efficient search for association rules”, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, p.99-107, Agosto 20-23, Boston, EUA.

Wettschereck, D. (2002), “A KDDSE-independent PMML Visualizer”, In Marko Bohanec, Dunja Mladenic, and Nada Lavrac, editors, *ECML/PKDD'02 workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, Agosto 2002.

Wong, P. C., P. Whitney e J. Thomas (1999), “Visualizing Association Rules for text Mining”, in IEEE Symposium on Information Visualization, San Francisco, CA.

Zakrzewicz, T. (2000), “Data Mining within DBMS functionality”; Proc. of PKDD 2000 Conference, Lyon, France.

Zhenjiang, H., W. Chin e M. Takeichi (2001), “Calculating a New Data Mining Algorithm for Market Basket Analysis”, Second International Workshop on Practical Aspects of Declarative Languages (PADL'00), Boston, Massachusetts, Janeiro 17-18, 2000. LNCS 1753, Springer Verlag. pp. 169-184.

Livros

Berry, M. J. A. e G. Linoff (1997), Data Mining Techniques: for Marketing, Sales and Customer Support, USA: John Wiley & Sons, inc.

Britt, J. e T. Duynstee (2000), Professional Visual Basic 6 XML, Birmingham, UK: Wrox Press Ltd.

Carli, J., J. Mason e R. V. Ramachandran (2000), Professional ASP Data Access, Birmingham, UK: Wrox Press Ltd.

McGrath, S. (1999), XML Aplicações Práticas, Rio de Janeiro, Brasil: Editora Campus.

Witten, I. e E. Frank (2000), Data Mining: Practical machine learning tools and techniques with Java implementations, USA: Morgan Kaufmann Publishers.

Sites

Adobe Systems Incorporated, Adobe SVG Viewer Download Area in http://www.adobe.com/svg/viewer/install/main.html

Data Mining Group (PMML development), Data Mining Group (PMML development), http://www.dmg.org/

Document Object Model (DOM), W3C Specifications in http://www.w3.org/DOM/

ECMA - Standardizing Information and Communication Systems, Standard ECMA-262 ECMAScript Language Specification in http://www.ecma.ch/ecma1/STAND/ECMA-262.HTM

Extensible Markup Language (XML), W3C Specifications and Drafts in http://www.w3.org/XML/

HyperText Markup Language (HTML), W3C Activity in http://www.w3.org/MarkUp/

Hypertext Transfer Protocol (HTTP), W3C Specifications in http://www.w3.org/Protocols/

Post Processing Environment for Association Rules (PEAR), in http://www.3pontos.com/pear/index.htm

Projecto Class, http://www.niaad.liacc.up.pt/~amjorge/Projectos/Class

Rulequest Research, in http://www.rulequest.com

Scalable Vector Graphics (SVG) 1.0 Specification, W3C Recommendation 04 September 2001 in http://www.w3.org/TR/SVG/

Wal-Mart Stores, Inc. at a Glance (Wal-Mart Store), http://www.walmartstores.com/

World Wide Web Consortium (W3C), http://www.w3c.org

Anexo

Modelo de regras de associação em PMML

Fonte: Ficheiro de *Logs* do *site* do Instituto Nacional de Estatística (http://www.ine.pt)

```
<?xml version="1.0"?>
<PMML version="1.1" >

<AssociationModel>
  <AssocInputStats
    numberOfTransactions = "99999"
    minimumSupport = "0.010000"
    minimumConfidence = "0.650000"
    numberOfItens ='8'
    numberOfItensets = '66'
    numberOfRules = '211'
   />

<AssocItem id='1' value='Estatisticas_Gerais'/>
<AssocItem id='2' value='Territorio_e_Ambiente'/>
<AssocItem id='3' value='Exportacoes_e_Importacoes'/>
<AssocItem id='4' value='Agricultura_e_Pescas'/>
<AssocItem id='5' value='Populacao_e_CondicoesSociais'/>
<AssocItem id='6' value='Economia_e_Financas'/>
<AssocItem id='7' value='Industria_e_Energia'/>
<AssocItem id='8' value='Comercio_Servicos_e_Turismo'/>

<AssocItemset id='1' support='0' numberOfItens='1'>
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='2' support='0' numberOfItens='2'>
<AssocItemRef itemRef='2' />
<AssocItemRef itemRef='9' />
</AssocItemset>
<AssocItemset id='3' support='0' numberOfItens='2'>
<AssocItemRef itemRef='2' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='4' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='2' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='5' support='0' numberOfItens='2'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='2' />
</AssocItemset>
<AssocItemset id='6' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='2' />
</AssocItemset>
<AssocItemset id='7' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='3' />
</AssocItemset>
<AssocItemset id='8' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='7' />
</AssocItemset>
<AssocItemset id='9' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='2' />
</AssocItemset>
<AssocItemset id='10' support='0' numberOfItens='3'>
```

```
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='3' />
</AssocItemset>
<AssocItemset id='11' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='12' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='2' />
</AssocItemset>
<AssocItemset id='13' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='3' />
</AssocItemset>
<AssocItemset id='14' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='15' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='7' />
</AssocItemset>
<AssocItemset id='16' support='0' numberOfItens='3'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='2' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='17' support='0' numberOfItens='2'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='2' />
</AssocItemset>
<AssocItemset id='18' support='0' numberOfItens='3'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='2' />
</AssocItemset>
<AssocItemset id='19' support='0' numberOfItens='3'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='3' />
</AssocItemset>
<AssocItemset id='20' support='0' numberOfItens='3'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='2' />
</AssocItemset>
<AssocItemset id='21' support='0' numberOfItens='3'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='3' />
</AssocItemset>
<AssocItemset id='22' support='0' numberOfItens='3'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='23' support='0' numberOfItens='3'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='7' />
</AssocItemset>
<AssocItemset id='24' support='0' numberOfItens='2'>
<AssocItemRef itemRef='4' />
<AssocItemRef itemRef='9' />
</AssocItemset>
<AssocItemset id='25' support='0' numberOfItens='2'>
<AssocItemRef itemRef='7' />
```

```
<AssocItemRef itemRef='2' />
</AssocItemset>
<AssocItemset id='26' support='0' numberOfItens='2'>
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='9' />
</AssocItemset>
<AssocItemset id='27' support='0' numberOfItens='2'>
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='28' support='0' numberOfItens='3'>
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='2' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='29' support='0' numberOfItens='2'>
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='2' />
</AssocItemset>
<AssocItemset id='30' support='0' numberOfItens='2'>
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='3' />
</AssocItemset>
<AssocItemset id='31' support='0' numberOfItens='2'>
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='32' support='0' numberOfItens='3'>
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='2' />
</AssocItemset>
<AssocItemset id='33' support='0' numberOfItens='3'>
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='3' />
</AssocItemset>
<AssocItemset id='34' support='0' numberOfItens='2'>
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='7' />
</AssocItemset>
<AssocItemset id='35' support='0' numberOfItens='1'>
<AssocItemRef itemRef='2' />
</AssocItemset>
<AssocItemset id='36' support='0' numberOfItens='2'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='37' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='38' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='39' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='8' />
</AssocItemset>
<AssocItemset id='40' support='0' numberOfItens='2'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='3' />
</AssocItemset>
<AssocItemset id='41' support='0' numberOfItens='2'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='42' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='7' />
```

```
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='43' support='0' numberOfItens='2'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='7' />
</AssocItemset>
<AssocItemset id='44' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='45' support='0' numberOfItens='2'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='46' support='0' numberOfItens='2'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='9' />
</AssocItemset>
<AssocItemset id='47' support='0' numberOfItens='2'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='4' />
</AssocItemset>
<AssocItemset id='48' support='0' numberOfItens='3'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='49' support='0' numberOfItens='3'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='50' support='0' numberOfItens='2'>
<AssocItemRef itemRef='4' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='51' support='0' numberOfItens='2'>
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='52' support='0' numberOfItens='2'>
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='53' support='0' numberOfItens='3'>
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='54' support='0' numberOfItens='1'>
<AssocItemRef itemRef='6' />
</AssocItemset>
<AssocItemset id='55' support='0' numberOfItens='3'>
<AssocItemRef itemRef='7' />
<AssocItemRef itemRef='2' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='56' support='0' numberOfItens='3'>
<AssocItemRef itemRef='8' />
<AssocItemRef itemRef='2' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='57' support='0' numberOfItens='1'>
<AssocItemRef itemRef='3' />
</AssocItemset>
<AssocItemset id='58' support='0' numberOfItens='3'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='2' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='59' support='0' numberOfItens='1'>
<AssocItemRef itemRef='4' />
</AssocItemset>
```

```
<AssocItemset id='60' support='0' numberOfItens='1'>
<AssocItemRef itemRef='7' />
</AssocItemset>
<AssocItemset id='61' support='0' numberOfItens='2'>
<AssocItemRef itemRef='2' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='62' support='0' numberOfItens='3'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='2' />
<AssocItemRef itemRef='1' />
</AssocItemset>
<AssocItemset id='63' support='0' numberOfItens='2'>
<AssocItemRef itemRef='5' />
<AssocItemRef itemRef='8' />
</AssocItemset>
<AssocItemset id='64' support='0' numberOfItens='2'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='9' />
</AssocItemset>
<AssocItemset id='65' support='0' numberOfItens='1'>
<AssocItemRef itemRef='8' />
</AssocItemset>
<AssocItemset id='66' support='0' numberOfItens='2'>
<AssocItemRef itemRef='6' />
<AssocItemRef itemRef='7' />
</AssocItemset>

<AssocRule support='0.032' confidence='0.654' antecedent='1' consequent='2' />
<AssocRule support='0.058' confidence='0.489' antecedent='1' consequent='3' />
<AssocRule support='0.049' confidence='0.475' antecedent='1' consequent='4' />
<AssocRule support='0.161' confidence='0.405' antecedent='1' consequent='5' />
<AssocRule support='0.071' confidence='0.5' antecedent='1' consequent='6' />
<AssocRule support='0.05' confidence='0.463' antecedent='1' consequent='7' />
<AssocRule support='0.081' confidence='0.424' antecedent='1' consequent='8' />
<AssocRule support='0.053' confidence='0.628' antecedent='1' consequent='9' />
<AssocRule support='0.038' confidence='0.613' antecedent='1' consequent='10' />
<AssocRule support='0.034' confidence='0.536' antecedent='1' consequent='11' />
<AssocRule support='0.062' confidence='0.588' antecedent='1' consequent='12' />
<AssocRule support='0.045' confidence='0.514' antecedent='1' consequent='13' />
<AssocRule support='0.048' confidence='0.487' antecedent='1' consequent='14' />
<AssocRule support='0.076' confidence='0.484' antecedent='1' consequent='15' />
<AssocRule support='0.031' confidence='0.6' antecedent='1' consequent='16' />
<AssocRule support='0.077' confidence='0.492' antecedent='1' consequent='17' />
<AssocRule support='0.042' confidence='0.647' antecedent='1' consequent='18' />
<AssocRule support='0.038' confidence='0.581' antecedent='1' consequent='19' />
<AssocRule support='0.047' confidence='0.684' antecedent='1' consequent='20' />
<AssocRule support='0.044' confidence='0.528' antecedent='1' consequent='21' />
<AssocRule support='0.045' confidence='0.432' antecedent='1' consequent='22' />
<AssocRule support='0.07' confidence='0.474' antecedent='1' consequent='23' />
<AssocRule support='0.033' confidence='0.407' antecedent='1' consequent='24' />
<AssocRule support='0.06' confidence='0.612' antecedent='1' consequent='25' />
<AssocRule support='0.047' confidence='0.5' antecedent='1' consequent='26' />
<AssocRule support='0.048' confidence='0.436' antecedent='1' consequent='27' />
<AssocRule support='0.033' confidence='0.667' antecedent='1' consequent='28' />
<AssocRule support='0.071' confidence='0.603' antecedent='1' consequent='29' />
<AssocRule support='0.054' confidence='0.455' antecedent='1' consequent='30' />
<AssocRule support='0.069' confidence='0.411' antecedent='1' consequent='31' />
<AssocRule support='0.04' confidence='0.727' antecedent='1' consequent='32' />
<AssocRule support='0.036' confidence='0.621' antecedent='1' consequent='33' />
<AssocRule support='0.097' confidence='0.405' antecedent='1' consequent='34' />
<AssocRule support='0.141' confidence='0.539' antecedent='35' consequent='1' />
<AssocRule support='0.103' confidence='0.631' antecedent='35' consequent='36' />
<AssocRule support='0.047' confidence='0.763' antecedent='35' consequent='37' />
<AssocRule support='0.05' confidence='0.488' antecedent='35' consequent='7' />
<AssocRule support='0.048' confidence='0.59' antecedent='35' consequent='38' />
<AssocRule support='0.081' confidence='0.485' antecedent='35' consequent='8' />
<AssocRule support='0.097' confidence='0.456' antecedent='35' consequent='39' />
<AssocRule support='0.062' confidence='0.431' antecedent='35' consequent='40' />
<AssocRule support='0.099' confidence='0.494' antecedent='35' consequent='41' />
<AssocRule support='0.043' confidence='0.771' antecedent='35' consequent='42' />
<AssocRule support='0.038' confidence='0.548' antecedent='35' consequent='10' />
<AssocRule support='0.034' confidence='0.679' antecedent='35' consequent='11' />
<AssocRule support='0.127' confidence='0.413' antecedent='35' consequent='43' />
```

<AssocRule support='0.05' confidence='0.732' antecedent='35' consequent='44' />
<AssocRule support='0.045' confidence='0.541' antecedent='35' consequent='13' />
<AssocRule support='0.048' confidence='0.615' antecedent='35' consequent='14' />
<AssocRule support='0.076' confidence='0.516' antecedent='35' consequent='15' />
<AssocRule support='0.053' confidence='0.721' antecedent='35' consequent='45' />
<AssocRule support='0.062' confidence='0.431' antecedent='35' consequent='46' />
<AssocRule support='0.069' confidence='0.446' antecedent='35' consequent='47' />
<AssocRule support='0.036' confidence='0.759' antecedent='35' consequent='48' />
<AssocRule support='0.038' confidence='0.548' antecedent='35' consequent='19' />
<AssocRule support='0.043' confidence='0.743' antecedent='35' consequent='49' />
<AssocRule support='0.044' confidence='0.528' antecedent='35' consequent='21' />
<AssocRule support='0.045' confidence='0.459' antecedent='35' consequent='22' />
<AssocRule support='0.07' confidence='0.474' antecedent='35' consequent='23' />
<AssocRule support='0.038' confidence='0.742' antecedent='35' consequent='50' />
<AssocRule support='0.033' confidence='0.444' antecedent='35' consequent='24' />
<AssocRule support='0.048' confidence='0.769' antecedent='35' consequent='51' />
<AssocRule support='0.047' confidence='0.474' antecedent='35' consequent='26' />
<AssocRule support='0.048' confidence='0.513' antecedent='35' consequent='27' />
<AssocRule support='0.062' confidence='0.686' antecedent='35' consequent='52' />
<AssocRule support='0.054' confidence='0.477' antecedent='35' consequent='30' />
<AssocRule support='0.069' confidence='0.482' antecedent='35' consequent='31' />
<AssocRule support='0.039' confidence='0.75' antecedent='35' consequent='53' />
<AssocRule support='0.036' confidence='0.552' antecedent='35' consequent='33' />
<AssocRule support='0.097' confidence='0.418' antecedent='35' consequent='34' />
<AssocRule support='0.032' confidence='0.846' antecedent='54' consequent='2' />
<AssocRule support='0.062' confidence='0.804' antecedent='54' consequent='40' />
<AssocRule support='0.043' confidence='0.8' antecedent='54' consequent='42' />
<AssocRule support='0.053' confidence='0.744' antecedent='54' consequent='9' />
<AssocRule support='0.038' confidence='0.935' antecedent='54' consequent='10' />
<AssocRule support='0.034' confidence='0.679' antecedent='54' consequent='11' />
<AssocRule support='0.127' confidence='0.635' antecedent='54' consequent='43' />
<AssocRule support='0.05' confidence='0.756' antecedent='54' consequent='44' />
<AssocRule support='0.062' confidence='0.706' antecedent='54' consequent='12' />
<AssocRule support='0.045' confidence='0.892' antecedent='54' consequent='13' />
<AssocRule support='0.048' confidence='0.667' antecedent='54' consequent='14' />
<AssocRule support='0.076' confidence='0.806' antecedent='54' consequent='15' />
<AssocRule support='0.033' confidence='0.704' antecedent='54' consequent='24' />
<AssocRule support='0.048' confidence='0.744' antecedent='54' consequent='51' />
<AssocRule support='0.037' confidence='0.733' antecedent='54' consequent='55' />
<AssocRule support='0.06' confidence='0.694' antecedent='54' consequent='25' />
<AssocRule support='0.047' confidence='0.816' antecedent='54' consequent='26' />
<AssocRule support='0.062' confidence='0.686' antecedent='54' consequent='52' />
<AssocRule support='0.043' confidence='0.743' antecedent='54' consequent='56' />
<AssocRule support='0.033' confidence='0.63' antecedent='54' consequent='28' />
<AssocRule support='0.071' confidence='0.655' antecedent='54' consequent='29' />
<AssocRule support='0.054' confidence='0.818' antecedent='54' consequent='30' />
<AssocRule support='0.069' confidence='0.661' antecedent='54' consequent='31' />
<AssocRule support='0.039' confidence='0.844' antecedent='54' consequent='53' />
<AssocRule support='0.04' confidence='0.818' antecedent='54' consequent='32' />
<AssocRule support='0.036' confidence='0.862' antecedent='54' consequent='33' />
<AssocRule support='0.097' confidence='0.722' antecedent='54' consequent='34' />
<AssocRule support='0.047' confidence='0.5' antecedent='57' consequent='37' />
<AssocRule support='0.081' confidence='0.439' antecedent='57' consequent='8' />
<AssocRule support='0.097' confidence='0.418' antecedent='57' consequent='39' />
<AssocRule support='0.043' confidence='0.543' antecedent='57' consequent='42' />
<AssocRule support='0.034' confidence='0.5' antecedent='57' consequent='11' />
<AssocRule support='0.05' confidence='0.463' antecedent='57' consequent='44' />
<AssocRule support='0.076' confidence='0.435' antecedent='57' consequent='15' />
<AssocRule support='0.053' confidence='0.465' antecedent='57' consequent='45' />
<AssocRule support='0.038' confidence='0.516' antecedent='57' consequent='58' />
<AssocRule support='0.031' confidence='0.44' antecedent='57' consequent='16' />
<AssocRule support='0.036' confidence='0.621' antecedent='57' consequent='48' />
<AssocRule support='0.042' confidence='0.5' antecedent='57' consequent='18' />
<AssocRule support='0.043' confidence='0.543' antecedent='57' consequent='49' />
<AssocRule support='0.047' confidence='0.5' antecedent='57' consequent='20' />
<AssocRule support='0.045' confidence='0.432' antecedent='57' consequent='22' />
<AssocRule support='0.07' confidence='0.439' antecedent='57' consequent='23' />
<AssocRule support='0.048' confidence='0.487' antecedent='57' consequent='51' />
<AssocRule support='0.037' confidence='0.5' antecedent='57' consequent='55' />
<AssocRule support='0.048' confidence='0.41' antecedent='57' consequent='27' />
<AssocRule support='0.043' confidence='0.457' antecedent='57' consequent='56' />
<AssocRule support='0.033' confidence='0.444' antecedent='57' consequent='28' />
<AssocRule support='0.039' confidence='0.562' antecedent='57' consequent='53' />
<AssocRule support='0.04' confidence='0.485' antecedent='57' consequent='32' />

<AssocRule support='0.032' confidence='0.462' antecedent='59' consequent='2' />
<AssocRule support='0.043' confidence='0.429' antecedent='59' consequent='42' />
<AssocRule support='0.053' confidence='0.442' antecedent='59' consequent='9' />
<AssocRule support='0.038' confidence='0.452' antecedent='59' consequent='10' />
<AssocRule support='0.05' confidence='0.463' antecedent='59' consequent='44' />
<AssocRule support='0.062' confidence='0.471' antecedent='59' consequent='12' />
<AssocRule support='0.045' confidence='0.405' antecedent='59' consequent='13' />
<AssocRule support='0.053' confidence='0.419' antecedent='59' consequent='45' />
<AssocRule support='0.038' confidence='0.484' antecedent='59' consequent='58' />
<AssocRule support='0.036' confidence='0.448' antecedent='59' consequent='48' />
<AssocRule support='0.042' confidence='0.412' antecedent='59' consequent='18' />
<AssocRule support='0.038' confidence='0.419' antecedent='59' consequent='19' />
<AssocRule support='0.043' confidence='0.457' antecedent='59' consequent='49' />
<AssocRule support='0.047' confidence='0.447' antecedent='59' consequent='20' />
<AssocRule support='0.044' confidence='0.444' antecedent='59' consequent='21' />
<AssocRule support='0.048' confidence='0.436' antecedent='59' consequent='51' />
<AssocRule support='0.037' confidence='0.5' antecedent='59' consequent='55' />
<AssocRule support='0.06' confidence='0.408' antecedent='59' consequent='25' />
<AssocRule support='0.047' confidence='0.421' antecedent='59' consequent='26' />
<AssocRule support='0.062' confidence='0.451' antecedent='59' consequent='52' />
<AssocRule support='0.043' confidence='0.514' antecedent='59' consequent='56' />
<AssocRule support='0.071' confidence='0.466' antecedent='59' consequent='29' />
<AssocRule support='0.054' confidence='0.432' antecedent='59' consequent='30' />
<AssocRule support='0.039' confidence='0.5' antecedent='59' consequent='53' />
<AssocRule support='0.04' confidence='0.485' antecedent='59' consequent='32' />
<AssocRule support='0.036' confidence='0.448' antecedent='59' consequent='33' />
<AssocRule support='0.076' confidence='0.484' antecedent='60' consequent='61' />
<AssocRule support='0.032' confidence='0.692' antecedent='60' consequent='2' />
<AssocRule support='0.065' confidence='0.509' antecedent='60' consequent='62' />
<AssocRule support='0.049' confidence='0.475' antecedent='60' consequent='4' />
<AssocRule support='0.047' confidence='0.737' antecedent='60' consequent='37' />
<AssocRule support='0.071' confidence='0.552' antecedent='60' consequent='6' />
<AssocRule support='0.05' confidence='0.707' antecedent='60' consequent='7' />
<AssocRule support='0.048' confidence='0.487' antecedent='60' consequent='38' />
<AssocRule support='0.097' confidence='0.633' antecedent='60' consequent='39' />
<AssocRule support='0.062' confidence='0.608' antecedent='60' consequent='40' />
<AssocRule support='0.05' confidence='0.732' antecedent='60' consequent='44' />
<AssocRule support='0.062' confidence='0.627' antecedent='60' consequent='12' />
<AssocRule support='0.045' confidence='0.73' antecedent='60' consequent='13' />
<AssocRule support='0.048' confidence='0.538' antecedent='60' consequent='14' />
<AssocRule support='0.158' confidence='0.481' antecedent='60' consequent='63' />
<AssocRule support='0.053' confidence='0.674' antecedent='60' consequent='45' />
<AssocRule support='0.038' confidence='0.71' antecedent='60' consequent='58' />
<AssocRule support='0.031' confidence='0.56' antecedent='60' consequent='16' />
<AssocRule support='0.077' confidence='0.54' antecedent='60' consequent='17' />
<AssocRule support='0.062' confidence='0.608' antecedent='60' consequent='46' />
<AssocRule support='0.043' confidence='0.771' antecedent='60' consequent='49' />
<AssocRule support='0.047' confidence='0.711' antecedent='60' consequent='20' />
<AssocRule support='0.044' confidence='0.694' antecedent='60' consequent='21' />
<AssocRule support='0.045' confidence='0.514' antecedent='60' consequent='22' />
<AssocRule support='0.127' confidence='0.548' antecedent='60' consequent='64' />
<AssocRule support='0.038' confidence='0.548' antecedent='60' consequent='50' />
<AssocRule support='0.033' confidence='0.593' antecedent='60' consequent='24' />
<AssocRule support='0.062' confidence='0.627' antecedent='60' consequent='52' />
<AssocRule support='0.043' confidence='0.686' antecedent='60' consequent='56' />
<AssocRule support='0.033' confidence='0.593' antecedent='60' consequent='28' />
<AssocRule support='0.071' confidence='0.569' antecedent='60' consequent='29' />
<AssocRule support='0.054' confidence='0.659' antecedent='60' consequent='30' />
<AssocRule support='0.076' confidence='0.565' antecedent='65' consequent='61' />
<AssocRule support='0.032' confidence='0.808' antecedent='65' consequent='2' />
<AssocRule support='0.058' confidence='0.574' antecedent='65' consequent='3' />
<AssocRule support='0.065' confidence='0.566' antecedent='65' consequent='62' />
<AssocRule support='0.049' confidence='0.6' antecedent='65' consequent='4' />
<AssocRule support='0.047' confidence='0.816' antecedent='65' consequent='37' />
<AssocRule support='0.071' confidence='0.621' antecedent='65' consequent='6' />
<AssocRule support='0.05' confidence='0.805' antecedent='65' consequent='7' />
<AssocRule support='0.048' confidence='0.667' antecedent='65' consequent='38' />
<AssocRule support='0.081' confidence='0.758' antecedent='65' consequent='8' />
<AssocRule support='0.062' confidence='0.725' antecedent='65' consequent='40' />
<AssocRule support='0.043' confidence='0.857' antecedent='65' consequent='42' />
<AssocRule support='0.053' confidence='0.744' antecedent='65' consequent='9' />
<AssocRule support='0.038' confidence='0.871' antecedent='65' consequent='10' />
<AssocRule support='0.034' confidence='0.75' antecedent='65' consequent='11' />
<AssocRule support='0.127' confidence='0.596' antecedent='65' consequent='43' />

```
<AssocRule support='0.053' confidence='0.814' antecedent='65' consequent='45' />
<AssocRule support='0.038' confidence='0.839' antecedent='65' consequent='58' />
<AssocRule support='0.031' confidence='0.68' antecedent='65' consequent='16' />
<AssocRule support='0.077' confidence='0.603' antecedent='65' consequent='17' />
<AssocRule support='0.062' confidence='0.706' antecedent='65' consequent='46' />
<AssocRule support='0.069' confidence='0.661' antecedent='65' consequent='47' />
<AssocRule support='0.036' confidence='0.931' antecedent='65' consequent='48' />
<AssocRule support='0.042' confidence='0.794' antecedent='65' consequent='18' />
<AssocRule support='0.038' confidence='0.806' antecedent='65' consequent='19' />
<AssocRule support='0.11' confidence='0.633' antecedent='65' consequent='66' />
<AssocRule support='0.038' confidence='0.742' antecedent='65' consequent='50' />
<AssocRule support='0.033' confidence='0.704' antecedent='65' consequent='24' />
<AssocRule support='0.048' confidence='0.821' antecedent='65' consequent='51' />
<AssocRule support='0.037' confidence='0.8' antecedent='65' consequent='55' />
<AssocRule support='0.06' confidence='0.673' antecedent='65' consequent='25' />
<AssocRule support='0.047' confidence='0.763' antecedent='65' consequent='26' />
<AssocRule support='0.048' confidence='0.615' antecedent='65' consequent='27' />
</AssociationModel>
</PMML>
```