Universidade Federal do Rio de Janeiro
Instituto de Matemática
Núcleo de Computação Eletrônica

Kelli de Faria Cordeiro

# *Cubing while Mining*: Ambiente Analítico para Apoio ao Processo de Exploração de Regras de Associação

Rio de Janeiro
2005

Kelli de Faria Cordeiro

# *Cubing while Mining*: Ambiente Analítico para Apoio ao Processo de Exploração de Regras de Associação

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática (Sistemas de Informação), Instituto de Matemática/Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática (Sistemas de Informação)

Pedro Manoel da Silveira, Ph.D.

Rio de Janeiro
2005

Cordeiro, Kelli de Faria

*Cubing while Mining*: Ambiente Analítico para Apoio ao Processo de Exploração de Regras de Associação / Kelli de Faria Cordeiro – Rio de Janeiro, 2005.

100f.: il.

Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro – UFRJ, Instituto de Matemática – IM, Núcleo de Computação Eletrônica – NCE, 2005.

Orientador: Pedro Manoel da Silveira

1. Processo de Descoberta do Conhecimento. 2. Mineração de Dados. 3. Data Warehouse.

# *Cubing while Mining*: Ambiente Analítico para Apoio ao Processo de Exploração de Regras de Associação

Kelli de Faria Cordeiro

Dissertação submetida ao corpo docente do Instituto de Matemática – Universidade Federal do Rio de Janeiro – UFRJ, como parte dos requisitos necessários à obtenção do grau de Mestre.

Aprovada por:

Orientador
Pedro Manoel da Silveira, Ph.D., IM/NCE - UFRJ

Maria Luiza Machado Campos, Ph.D., IM - UFRJ

Eber Assis Schmitz, Ph.D., IM/NCE - UFRJ

Alexandre Plastino de Carvalho, D.Sc., IC - UFF

# DEDICATÓRIA

*Dedico a **Deus***
*por ter me iluminado, guiado e colocado as pessoas certas no meu caminho.*

# AGRADECIMENTOS

*A única maneira de iniciar os agradecimentos é pela pessoa que despertou em mim o sonho de fazer um mestrado, a **Glenda**. Eu achava que era algo que não poderia alcançar, então nem fazia parte dos meus sonhos. Foi ela quem primeiro acreditou que eu poderia fazer o mestrado, antes de mim mesma! Seu entusiasmo e confiança me trouxeram até aqui. E foi assim que comecei a trilhar esse caminho, seguindo seus passos, seu exemplo, como tenho feito até hoje. Glenda, obrigada por ter acreditado em mim e ter me levado até esse mundo de conhecimento, crescimento e amadurecimento! Devo tudo isso a você!*

*Por meio dela vieram outras pessoas, tão entusiasmadas quanto, que acreditaram nesse sonho, até então pra mim impossível. Uma dessas pessoas é o **George Hamilton**, que não mediu esforços na minha condução pelo processo seletivo.*

*Foi um longo processo seletivo com mais de um ano de intensas avaliações. E começou com a Professora **Maria Luiza**, uma pessoa especial, que me recebeu de uma maneira muito "especial", compreendendo as limitações que eu trazia comigo. Agradeço imensamente sua generosidade, confiança e orientações nos primeiros passos no caminho da pesquisa e ao longo de todo o mestrado.*

*Durante esse longo processo seletivo e durante todo o curso, tive o privilégio de estar ao lado do **Moriya**, um grande amigo! Uma pessoa rara, com tranqüilidade, sensatez, inteligência e um grande companheirismo. Foi um privilégio tê-lo como parceiro em quase todos os trabalhos em grupo e, com isso, ter aprendido tantas coisas novas.*

*E assim, vieram outros colegas que muito colaboraram para o meu aprendizado, com dicas, ajudas e amizade, entre eles: **Simone Garcia**, **Linair**, **Sérgio Serra**, **Alissandra**, **Renata Chaomey**, **Eric Praxedes**, **Wilson**, **Pablo**, **Cássio**, **Luciana**, **Fabrício** e **Célia Seabra**.*

*Além dos colegas, vieram vários professores, que pacientemente me ensinaram muito! Dentre eles, está a Professora **Lígia**, que com seus ensinamentos ampliou minha maneira de ver e pensar cientificamente, mudando bastante a minha visão sobre pesquisa científica. Além de ser uma das melhores professoras que já tive e de ter uma das melhores aulas do mestrado, é uma pessoa muito especial. Sua sensibilidade nos ensina a pensar, a ter consciência! Obrigada, Lígia, pelos óculos azul.*

*Um elemento fundamental do meu mestrado foi o meu orientador, o Professor **Pedro Manoel**, que, ao longo dessa jornada, com seus ensinamentos, incentivos e maneira prática de ser, ajudou-me a ampliar meus conhecimentos e a desenvolver esta dissertação. Agora, já no final desta etapa, refletindo sobre tudo o que aconteceu, vêm à minha memória todos os seus incentivos às iniciativas empolgadas de um aluno em início de pesquisa. Ele deixou que eu fosse percebendo por mim mesma a inviabilidade das iniciativas. E foi assim que me conduziu, dando espaço para que eu percebesse o melhor caminho a seguir, estando presente nos momentos decisivos. Muito obrigada Professor Pedro, pela paciência, compreensão e imprescindíveis orientações!*

*Agradeço à Marinha do Brasil por viabilizar a minha participação no mestrado. Em especial, agradeço ao Almirante **Ostwald**, ao Almirante **Barros**, ao Almirante **Vasquez***

*Gomes, ao Comandante Allevato e ao Comandante Rodrigues Neto. Espero poder retribuir com um trabalho técnico profissional à altura de suas expectativas.*

*Agradeço, também, ao meu chefe de divisão, Comandante Barros, uma pessoa muito sábia e sensata, que lidera seus subordinados com muita propriedade. É um constante motivador ao estudo e ao aprimoramento profissional. Além de motivar, ele busca os meios para viabilizar os estudos. Essa busca nem sempre é fácil, mas sua determinação faz com que os meios cheguem até nós. E tudo isso nos leva a uma boa prática profissional, com trabalhos de qualidade e alto nível. Para mim, sua maneira democrática de liderar é um grande exemplo.*

*Agradeço, ainda, a outras pessoas da Marinha, que me deram um suporte importante e necessário para a realização deste trabalho: André Vitor, meu melhor amigo, presente em todas as horas, com sua leveza e maneira prática de ser, e conselhos sempre pertinentes; Queiroz, com sua compreensão, constante incentivo e orientação sobre os procedimentos internos; Luiz Fraga, sempre presente e disponível para ajudar; Parêsqui, que mesmo querendo que eu me dedique mais ao esporte, me incentiva com a sua admiração pelo meu trabalho; e os demais colegas de trabalho, que tiveram participação constante: Arnaldo, Márcia Aboim, Denise, Rosana e Sidney.*

*Agradeço às amigas que conheci na Marinha, que compreenderam a minha ausência nos diversos eventos que não pude comparecer: Camila, Sheyla, Simone Aragão, Elisabeth Medeiros, Márcia Tenório, Paulinha, Estrela, Tinoco, Rachel e Elisabete.*

*No âmbito familiar, agradeço a todos pelo incentivo ao estudo, pela admiração e pela compreensão à minha ausência. Em especial ao meu pai Geraldo, à minha mãe Glória e à minha irmã Hebe.*

*Agradeço à pessoa que viu, sentiu, acompanhou e vibrou intensamente com toda essa conquista, o Denys, meu marido. Nossa união aconteceu no meio desse caminho não muito tranqüilo, e ele soube compreender todas as minhas faltas e falhas. Sua admiração e orgulho foram os motivadores nas noites em claro, nas madrugadas e nos dias de muito estudo e trabalho. Sua compreensão inquestionável pelos tantos passeios que deixamos de fazer, deu a tranqüilidade necessária para que eu me dedicasse ao estudo. Denys, essa conquista também é sua!*

*No meio desse longo caminho pude fazer algumas novas amizades, que não se importaram com a minha limitação de tempo. Dentre elas está a Kilza, uma pessoa admirável pelo companheirismo, inteligência, sensibilidade e percepção, que me ajudou a compreender as desilusões e a me acalmar nos momentos de muita angústia.*

*Outra pessoa que esteve muito próxima durante esse período foi a Vaninha. Graças às facilidades da internet, esteve em muitos momentos diante do computador, incentivando e dando muitas orientações e dicas. Com sua força contagiante, não permitiu que eu esmorecesse nos momentos de cansaço e serviu de grande exemplo na busca pela qualidade das produções acadêmicas.*

*Agradeço a participação de todos os membros da banca examinadora: Professor* ***Pedro Manoel****, Professora* ***Maria Luiza****, Professor* ***Eber****, e nosso convidado de fora do programa, o Professor* ***Plastino****.*

*Agradeço à* ***Renata Lavôr*** *por ter cedido o código fonte da implementação da sua dissertação, pela atenção e apoio no início deste trabalho. E ao* ***Rodrigo Salvador*** *pelas suas pesquisas que estiveram presentes na base deste trabalho.*

*Agradeço às recomendações e incentivo ao estudo dos professores:* ***Blascheck****,* ***Vicente Fernandes****,* ***Pereira Neto****,* ***Jucele*** *e* ***Vasconcellos****.*

*A ordem desta lista de agradecimentos não tem grau de importância, todos tiverem um papel fundamental, independente do quanto contribuíram. Saber que podemos contar com os que estão à nossa volta nos dá um enorme conforto. E, a todos vocês, mesmo os que não foram citados, minha eterna gratidão!*

***Kelli***

# RESUMO

CORDEIRO, Kelli de Faria. ***Cubing while Mining*: Ambiente Analítico para Apoio ao Processo de Exploração de Regras de Associação**. Rio de Janeiro, 2005. Dissertação (Mestrado em Informática) - Instituto de Matemática/Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2005.

O processo de mineração de regras de associação, tipicamente, gera muitas regras, nem sempre interessantes ou úteis e de manipulação complexa. Uma das formas para contornar esse problema é permitir uma interação freqüente entre o usuário e o sistema durante a mineração de regras de associação, possibilitando análises dos resultados parciais. A proposta deste trabalho é a definição de um ambiente analítico de apoio ao processo de mineração, que explore informações pertinentes ao próprio processo, de modo a orientar e permitir a interferência do usuário visando à obtenção das regras mais interessantes. Nesse ambiente, um *data mart* é alimentado com as regras de associação finais e candidatas, geradas a cada ciclo do algoritmo *Apriori*, além de diversas medidas de interesse. Dessa forma, é possível avaliar os resultados parciais da mineração, calibrar e redefinir as medidas de interesse e realimentar o algoritmo. O processo analítico para exploração de regras, a arquitetura do ambiente e o modelo de dados multidimensional das regras de associação são as principais contribuições deste trabalho.

# ABSTRACT

CORDEIRO, Kelli de Faria. ***Cubing while Mining*: Ambiente Analítico para Apoio ao Processo de Exploração de Regras de Associação**. Rio de Janeiro, 2005. Dissertação (Mestrado em Informática) - Instituto de Matemática/Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2005.

*The mining of association rules usually generates large numbers of rules, many of them not usefull at all, and poses a significant manipulation difficulty. By allowing a frequent interaction between the user and the system during the association rule mining we believe one can minimize this problem. This work proposes the definition of an analytical environment to support the mining process by exploring relevant data regarding the process itself, guiding the user towards the most interesting rules. In the proposed environment, a datamart is loaded with the association rules generated in each cycle of the Apriori algorithm, as well as the control measures employed. This way, it becomes possible to evaluate parcial results, adjust and redefine the interesting measures thus reloading them in the algorithm. The analytical process for rule exploration, the environment architecture and the multidimensional data model of the association rules are the main contributions of this work.*

# LISTA DE ILUSTRAÇÕES

# LISTA DE TABELAS

# LISTA DE SIGLAS

| | |
|---|---|
| DW | *Data Warehouse* |
| WWW | *World Wide Web* |
| OLAP | *On-Line Analytical Processing* |
| KDD | *Knowledge Discovery in Databases* |
| DIC | *Dynamic Itemset Counting* |
| OLAM | *On-Line Analytical Mining* |

# SUMÁRIO

# 1 INTRODUÇÃO

Mineração de regra de associação é uma tarefa de descoberta de padrões interessantes em um conjunto de itens, onde o dado pode estar armazenado em bancos de dados transacionais (relacionais, orientados-a-objeto e objeto-relacionais), bancos de dados multidimensionais (*data warehouses* - DW) ou outros repositórios de informação (bancos de dados textuais, multimídia e dados da *World Wide Web* - WWW). A mineração de dados faz parte de um campo interdisciplinar, originado de áreas como sistemas de banco de dados, *data warehousing*, estatística, *machine learning*, visualização e recuperação de informações, e computação de alto desempenho (HAN e KAMBER, 2001).

A mineração de dados utiliza ferramentas para revelar padrões de dados, como as associações, onde regras são geradas para permitir o entendimento do dado. A descoberta de relacionamentos de associação é realizada por meio da aplicação de algoritmos em conjuntos de itens de dados (HAN e KAMBER, 2001).

O ponto forte da mineração de regras de associação é sua completude, por encontrar todas as associações nos dados que satisfazem parâmetros especificados pelo usuário. Esse ponto forte, entretanto, pode acarretar o seguinte problema: um processo de mineração de regras de associação, a depender das medidas inicialmente estabelecidas pelo usuário, normalmente gera resultados cuja análise e manipulação tornam-se complexas. Essa complexidade decorre do número de regras, especialmente quando os dados de um conjunto são altamente correlacionados. Além disso, as regras geradas nem sempre são interessantes ou úteis, e regras interessantes podem não constar no resultado final.

Várias técnicas foram propostas para impedir a sobrecarga do usuário com o excesso de padrões, como OLAP *Mining* (HAN, 1997), direcionamento e redução do número de regras (LAVÔR, 2003), DWARF (MONTEIRO et al., 2003) e técnicas de visualização

(BRUZZESE e BUONO, 2004). Entretanto, essas técnicas ainda não solucionam de forma satisfatória o problema do excesso de regras geradas, pois regras não interessantes podem estar sendo geradas e regras interessantes podem estar sendo excluídas do resultado final. Portanto, encontrar relacionamentos relevantes entre os muitos padrões descobertos apresenta alguns desafios: como avaliar o que é útil, como selecionar o que interessa ou não, como gerar um número menor de regras e revelar mais conhecimentos novos e úteis.

Uma das formas para contornar esse problema é permitir uma interação freqüente entre o usuário e o sistema durante a realização da mineração de regras de associação, possibilitando análises dos resultados parciais. Para isso, é necessário um ambiente que permita ao usuário analisar e interagir com o processo de geração de regras.

Adotando a premissa de que a mineração de dados deve ser um processo centrado no elemento humano, onde a interferência do usuário permeia toda a mineração, este trabalho propõe uma arquitetura de exploração de regras, em um ambiente analítico, para apoiar tal processo com o propósito de facilitar e simplificar a análise dos resultados intermediários e finais.

No contexto desta pesquisa, um padrão de dado é interessante caso possua as seguintes características (Han e Kamber, 2001, p.27): "(i) fácil de entender por humanos; (ii) válido; (iii) potencialmente útil; e (iv) novo. Um padrão é interessante, também, se valida uma hipótese que o usuário deseja confirmar". Um ambiente analítico é definido como um ambiente de apoio ao processo de tomada de decisão caracterizado por: tipo de processamento *On-Line Analytical Processing* (OLAP); "pequeno" número de consultas "variáveis"; necessidade de visualização dos dados sob diferentes perspectivas; operações de agregação e cruzamento; relevância dos dados históricos; e necessidade de consistência. Um ambiente analítico possui, ainda, os seguintes requisitos: flexibilidade, facilidade de navegação;

consultas complexas, não antecipadas; gerenciamento de grande volume de dados; e necessidade de examinar o dado em diferentes níveis de detalhe (CAMPOS, 2004).

Na arquitetura proposta, as funcionalidades OLAP se integram ao processo de mineração, como inicialmente visto em Han (1997). Os dados originais, o resultado final do processo e o resultado de cada um dos ciclos do algoritmo *Apriori* (AGRAWAL e SRIKANT, 1994) são disponibilizados em um ambiente analítico, onde uma ferramenta OLAP é utilizada, pelo usuário, para análise dos dados sob diferentes perspectivas. Um modelo de dados multidimensional da memória do processo de exploração de regras de associação é projetado para permitir a realização dessa análise. Assim, o usuário pode visualizar o caminho que o algoritmo está percorrendo, identificar e avaliar as informações fornecidas, e redirecionar a mineração, quando necessário. A nova direção é definida por meio das medidas de interesse e dos parâmetros de entrada que podem realimentar o algoritmo a cada ciclo. Ao final do processo, o usuário poderá, também, selecionar as regras mais pertinentes e confrontá-las com os dados originais.

Acredita-se que a definição do ambiente analítico, em uma arquitetura com implementação viável, apoiará a mineração de regras de associação mais interessantes, revelando conhecimento novo e útil. Acredita-se, também, que a interferência do usuário no processo de mineração, redefinindo as medidas de interesse e os parâmetros de entrada durante a exploração, contribuirá para a redução da quantidade de regras do resultado final.

Assim sendo, a presente pesquisa propõe um ambiente analítico que apóia o analista de negócios na escolha do que é útil, e propõe, também, um processo analítico de exploração que permite a interferência do usuário para seleção das regras mais pertinentes, por meio da redefinição das medidas de interesse e dos parâmetros de entrada a cada ciclo do algoritmo *Apriori*. Com isso, espera-se contribuir para a redução do número de elementos do conjunto

de regras no final do processo de mineração. Presume-se que, nesse conjunto reduzido, estejam contidas as regras mais interessantes que revelam novos conhecimentos.

As contribuições esperadas neste estudo são as seguintes:

- Definição de uma arquitetura com todos os elementos necessários para um ambiente analítico de exploração de regras de associação;
- Definição de um processo analítico de exploração de regras de associação, com mecanismos que permitam a interferência do usuário no processo e direcionem a mineração de regras para as mais interessantes;
- Definição de mais um passo no processo de descoberta do conhecimento, para a análise dos resultados intermediários da mineração de regras de associação; e
- Definição do modelo de dados multidimensional do processo de exploração de regras de associação, para guardar a memória da mineração, os valores dos parâmetros utilizados e os resultados parciais e finais obtidos.

Esta dissertação está organizada da seguinte maneira: o capítulo 2 introduz os principais conceitos relativos à descoberta do conhecimento em banco de dados, abordando *data warehousing* e mineração de dados; o capítulo 3 descreve, formalmente, as regras de associação, apresenta algumas medidas de interesse e algoritmos utilizados para minerar regras de associação, e detalha o funcionamento do algoritmo *Apriori*; o capítulo 4 expõe as principais abordagens encontradas na literatura para extração e manipulação do excesso de regras de associação; uma arquitetura de exploração de regras de associação em um ambiente analítico é proposta no capítulo 5 e sua utilização em um processo interativo é discutida no capítulo 6; um exemplo de aplicação da arquitetura proposta é apresentado no capítulo 7 e o capítulo 8 conclui o trabalho com algumas considerações finais e trabalhos futuros.

# 2 DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS

O processo de descoberta do conhecimento, segundo Han e Kamber (2001), consiste em uma seqüência iterativa dos seguintes passos (Figura 1):

- Limpeza de dados: para remover ruídos e dados inconsistentes;
- Integração de dados: onde múltiplas fontes de dados podem ser combinadas;
- Seleção de dados: onde dados relevantes para a tarefa de análise são recuperados do banco de dados;
- Transformação de dados: onde dados são transformados ou consolidados em formulários apropriados para mineração;
- Mineração de dados: como um processo essencial onde métodos inteligentes são aplicados com o objetivo de extrair padrões de dados;
- Avaliação de padrões: para identificar padrões que representam conhecimento baseado em alguma medida de interesse; e
- Apresentação do conhecimento: onde as técnicas de visualização e representação do conhecimento são usadas para apresentar o conhecimento minerado para o usuário.

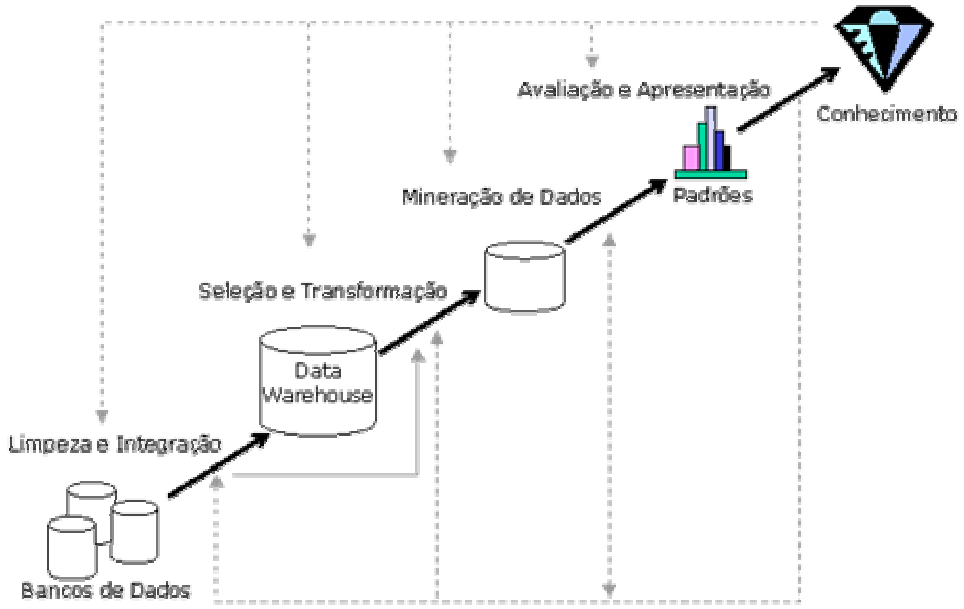



**Figura 1: Mineração de dados como um passo no processo de descoberta do conhecimento**
**Fonte: (HAN e KAMBER, 2001)**

As próximas seções abordam os principais conceitos relativos a duas das principais etapas do processo de descoberta do conhecimento: *data warehousing* e mineração de dados.

## 2.1 DATA WAREHOUSING

Inmon, em 1997, definiu *data warehouse* como uma coleção de dados orientada a assunto, integrada, variante no tempo e não-volátil, que apóia a gerência do processo de tomada de decisão e são mantidos separados da base de dados operacional. *Data warehousing* é o processo de construção e utilização de *data warehouses* (HAN e KAMBER, 2001).

Um *data warehouse* é orientado a assunto, pois é organizado em torno de um assunto e oferece uma visão simples e concisa sobre um assunto em particular, excluindo dados que não são úteis para o processo de tomada de decisão; é integrado, pois, normalmente, é construído

pela integração de fontes múltiplas de dados heterogêneos; é histórico, pois o dado é armazenado para oferecer uma perspectiva histórica, onde cada estrutura chave em um *data warehouse* contém um elemento de tempo explícito ou implícito; é não-volátil, pois não ocorre a operação de atualização de dados.

Fisicamente, um *data warehouse* é mantido separadamente de um ambiente operacional para que suas consultas analíticas não afetem o desempenho do ambiente operacional, pois normalmente essas consultas agregam muitos dados. Um *data warehouse* necessita, apenas, da operação de carga inicial, para extrair os dados do ambiente operacional, e da operação de acesso, para consultar os dados consolidados e integrados. Portanto, não necessita do processamento de transações de recuperação, nem de mecanismos de controle de concorrência para as operações de exclusão e alteração.

### 2.1.1 REPOSITÓRIO DE METADADOS

Metadados são dados sobre dados. No *data warehouse*, metadados são os dados que definem os objetos do repositório e desempenham um papel diferente dos outros dados do *data warehouse*. Os metadados são importantes para o *data warehouse* por sua utilização: (i) como um diretório para ajudar o analista de sistemas de suporte à decisão a localizar o conteúdo do *data warehouse*, (ii) como um guia para o mapeamento do dado quando o mesmo é transformado e transportado do ambiente operacional para o ambiente do *data warehouse*, e (iii) como um guia para os algoritmos de agregação entre o detalhe corrente e os dados menos agregados, e entre os dados menos agregados e os mais agregados. O metadado deve ser armazenado e gerenciado persistentemente, isto é, no disco rígido (HAN e KAMBER, 2001).

### 2.1.2 DATA MART

Em *data warehousing*, há uma distinção entre um *data warehouse* e um *data mart*. Um *data warehouse* coleta informações sobre assuntos que permeiam toda a organização e, portanto, tem um escopo corporativo. Um *data mart*, por outro lado, é um subconjunto de um *data warehouse* que tem o foco em um determinado assunto e, portanto, seu escopo é departamental.

De acordo com a fonte de dados do ambiente operacional, o *data mart* pode ser categorizado como independente ou dependente. Dependente, quando a fonte de dados é obtida diretamente de um *data warehouse* corporativo. Independente, quando a fonte de dados é capturada de um ou mais sistema transacional, de fornecedores de informações externos, ou de dados gerados localmente em um departamento, em particular, ou em uma área geográfica (HAN e KAMBER, 2001).

### 2.1.3 MODELO DE DADOS MULTIDIMENSIONAL

*Data warehouse* e ferramentas OLAP são baseadas em um modelo de dados multidimensional, que é, tipicamente, organizado em torno de um tema central. Esse tema é representado em uma tabela chamada “fato”. Fatos são medidas numéricas que permitem a análise do relacionamento entre as “dimensões”. Dimensões são as perspectivas ou entidades sobre as quais se deseja manter registro (HAN e KAMBER, 2001).

O modelo de dados multidimensional visualiza o dado no formato de um cubo, que é definido por dimensões e fatos. Um cubo de dados permite que o dado seja modelado e visualizado em múltiplas dimensões.

Os esquemas do modelo de dados multidimensionais podem ter as seguintes formas (KIMBALL et al., 1998):

- Esquema estrela (*star schema*): é constituído por uma grande tabela central (tabela fato) contendo a maior parte dos dados sem redundância, e por um conjunto de tabelas menores (tabelas dimensão), uma para cada dimensão;
- Esquema floco de neve (*snowflake*): é uma variação do esquema estrela, onde algumas tabelas dimensão são normalizadas, e assim o dado é dividido em tabelas adicionais; e
- Constelação de fatos (*facts constallation*): é constituído por múltiplas tabelas fato a serem compartilhadas entre as dimensões. Esse tipo de esquema pode ser visto como uma coleção de estrelas, e daí chamado de esquema galáctico ou constelação de fatos.

Para *data warehouses*, o esquema de constelação de fatos é comumente usado, visto que pode modelar assuntos múltiplos e inter-relacionados. Para *data marts*, o esquema floco de neve é comumente usado, visto que são orientados à modelagem de um único assunto (HAN e KAMBER, 2001).

### 2.1.4 PROCESSAMENTO ANALÍTICO

Um processamento analítico é considerado como a análise multidimensional dos dados de um *data warehouse*, com a utilização das operações básicas de uma ferramenta OLAP (HAN e KAMBER, 2001):

- *drill-down*: também conhecida como *roll-down*, ocorre quando o usuário aumenta o nível de detalhe da informação, diminuindo a granularidade, pois navega de um nível mais alto de agregação para um nível mais baixo de agregação, ou introduz novas dimensões (HAN e KAMBER, 2001);

- *drill-up*: também conhecida como *roll-up*, é o contrário da operação *drill-down* e ocorre quando o usuário aumenta a granularidade, diminuindo o nível de detalhamento da informação, pois navega para um nível hierárquico acima;
- *drill-across*: executa consultas que envolvem mais de uma tabela fato, solicitando os mesmos dados de uma outra tabela fato (KIMBALL, 2003);
- *drill-throught*: navega do nível mais baixo de um cubo de dados para as tabelas relacionais;
- *pivoting*: também conhecida como rotação, oferece uma visão alternativa do dado, obtida com a rotação de 90 graus do que está sendo visualizado. A visão é rotacionada no sentido horário sem qualquer reestruturação dos dados, a facilidade e a velocidade da execução dessa operação são exemplos de vantagem intrínseca da manipulação de uma matriz multidimensional; e
- *slice* e *dice*: também conhecidas como seleção e projeção, a operação de *slice* executa uma seleção em uma das dimensões do cubo resultando em um subcubo. A operação *dice* define um subcubo executando uma operação de projeção em uma ou mais dimensões.

Os sistemas de processamento analítico podem organizar e apresentar o dado em vários formatos, com o objetivo de acomodar as diversas necessidades dos usuários no processo de tomada de decisão. Exemplos desses formatos são tabelas ordenadas por determinadas colunas e gráficos de barra e pizza.

## 2.2 MINERAÇÃO DE DADOS

A mineração de dados tem despertado o interesse da indústria da informação devido à grande disponibilidade de dados e à necessidade de transformá-los em informações úteis e em conhecimento, que poderão ser usados em aplicações que abrangem desde gestão de negócios, controle da produção e análise de mercado, até projeto de engenharia e exploração científica.

A mineração de dados pode ser vista como o resultado da evolução natural da tecnologia da informação e refere-se à extração ou à mineração de conhecimento de um grande volume de dados.

As funcionalidades da mineração de dados são usadas para especificar os tipos de padrões a serem encontrados nas tarefas de mineração. Em geral, as tarefas de mineração de dados podem ser classificadas em duas categorias: descritiva e preditiva. Tarefas de mineração descritiva caracterizam as propriedades do dado em um banco de dados. Tarefas de mineração preditiva executam inferências nos dados correntes com o intuito de fazer predições (HAN e KAMBER, 2001).

### 2.2.1 MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO

A mineração de dados possui tarefas que visam à descoberta de padrões e tendências escondidas em conjuntos de dados. Uma das tarefas é a "associação", que estuda padrões de relacionamento entre itens de dados, gerando regras de associação (HAN e KAMBER, 2001). Um exemplo da tarefa de associação pode ser visto na livraria *Amazon* (*http://www.amazon.com*): ao pesquisar por um livro, a livraria apresenta o livro solicitado e sugestões de outros livros que foram comprados ou considerados, também, por outros consumidores. As sugestões são exibidas da seguinte forma: "*Customers who bought this book also bought:* (Clientes que compraram este livro também compraram:)" e "*Customers interested in (Data Mining: Concepts and Techniques) may also be interested in:* (Clientes interessados em (*Data Mining: Concepts and Techniques*) podem também estar interessados em:)".

O exemplo clássico de mineração de regras de associação é a análise de cesto de mercado (*market basket analysis*). Esse processo analisa os hábitos de compra de um consumidor encontrando associações entre os diferentes itens da sua cesta de compras. A

descoberta de tais associações pode ajudar o varejista a desenvolver estratégias de *marketing* adquirindo percepção de quais itens são freqüentemente adquiridos em conjunto pelos consumidores.

Mineração de regras de associação visa encontrar padrões freqüentes, associações, correlações ou estruturas causais entre um conjunto de itens em transações de banco de dados ou em outro repositório de informação.

### 2.2.2 ANÁLISE DE CORRELAÇÃO

A maioria dos algoritmos de mineração emprega o *framework* de suporte-confiança para dizer quais regras são interessantes para o usuário. Mesmo usando suporte e confiança mínima iniciais para ajudar a limpar ou excluir a exploração de regras desinteressantes, muitas regras que não são interessantes para o usuário ainda são produzidas (HAN e KAMBER, 2001).

Uma regra pode ser julgada interessante ou não tanto subjetivamente como objetivamente. Apenas o usuário pode julgar se uma dada regra é interessante ou não, e esse julgamento, sendo subjetivo, pode diferir de um usuário para outro. Entretanto, medidas de interesse objetivas, baseadas em estatísticas por trás dos dados, podem ser usadas como um passo em direção ao objetivo de eliminar regras desinteressantes da apresentação para o usuário (HAN e KAMBER, 2001).

Na mineração baseada em restrições, a mineração é executada sob o guia de vários tipos de restrições fornecidas pelo usuário:

- Restrições baseadas em conhecimento: especificam o tipo de conhecimento a ser minerado, como associação;
- Restrições de dados: especificam o conjunto de dados relevantes;

- Restrições dimensão/nível: especificam a dimensão dos dados, ou níveis da hierarquia conceitual a ser usada;
- Restrições de interesse: especificam medidas estatísticas de interesse da regra, como suporte e confiança; e
- Restrições de regra: especificam a forma da regra a ser minerada, e podem ser expressas: como metaregras (*templates* de regras), como o número máximo e mínimo de predicados que podem ocorrer em um antecedente ou conseqüente da regra, ou como relacionamentos entre atributos, valores de atributos e/ou agregações.

As metaregras permitem que os usuários especifiquem formas de regras que são interessantes. As formas das regras podem ser usadas como restrições para ajudar a melhorar a eficiência no processo de mineração. As metaregras podem ser baseadas na experiência do analista, nas expectativas ou na intuição relativas ao dado, ou automaticamente geradas baseadas no esquema do banco de dados (HAN e KAMBER, 2001).

# 3 REGRAS DE ASSOCIAÇÃO

As regras de associação podem ser vistas como um tipo de regra `se-então`: se uma pessoa compra o livro *Data Mining: Concepts and Techniques*, também compra *Data Mining: Introductory and Advanced Topics*. A diferença entre regras de associação e regras `se-então`, é uma probabilidade condicional. Se uma pessoa compra o livro A, há uma probabilidade de que também compre o livro B. Essa probabilidade condicional é conhecida, na literatura, como a medida de interesse **confiança** (AGRAWAL et al., 1993). Outra medida que é geralmente associada às regras de associação é o **suporte**: a fração de clientes que sustentam a regra, ou seja, o número de clientes que compraram todos os itens ocorridos na regra. Se há apenas um cliente que tenha comprado o livro A e B, então a regra de associação não é interessante (SIEBES e FEELDERS, 2003).

## 3.1 DESCRIÇÃO FORMAL

Formalmente, regras de associação são descritas assim:

> "seja $\Upsilon = I_1, I_2, ..., I_m$ um conjunto de atributos binários, chamados de itens. Seja $T$ um banco de dados de transações. Cada transação $t$ é representada como um vetor binário, com $t[k] = 1$ se $t$ possui o item $I_k$, e $t[k] = 0$, caso contrário. Existe uma tupla no banco de dados para cada transação. Seja $X$ um conjunto de alguns itens em $\Upsilon$. Dizemos que uma transação $t$ satisfaz $X$ se para todos os itens $I_k$ em $X$, $t[k] = 1$." (AGRAWAL et al., 1993)

Uma regra de associação pode ser entendida como uma implicação da forma $X \Rightarrow I_j$, onde $I_j$ é um item em $\Upsilon$ que não está presente em $X$, sendo $X$ chamado antecedente e $I_j$ conseqüente da regra. A regra $X \Rightarrow I_j$ é satisfeita no conjunto de transações $T$ com fator de confiança $0 \leq c \leq 1$, se e somente se pelo menos c% das transações em $T$ que satisfaçam $X$ também satisfizerem $I_j$ (SIEBES e FEELDERS, 2003). No contexto desta dissertação, a regra

de associação é mais genérica do que a definida em Agrawal et al. (1993), onde o conseqüente pode conter mais de um item, conforme definição de Agrawal e Srikant (1994) e descrito no item 3.5.1. Desta forma a regra de associação é entendida como uma implicação da forma $X \Rightarrow I$.

Dado um conjunto de transações *T*, é importante que sejam geradas todas as regras que satisfaçam certas restrições adicionais, como o suporte. Essa restrição refere-se ao número de transações em *T* que atende a uma regra. O suporte para a regra é definido como a fração de transações em *T* que satisfazem a união dos itens no conseqüente e antecedente da regra. Dados a confiança mínima e o suporte mínimo, os algoritmos processam todas as regras de associação que possuem confiança e suporte maior que o especificado (SIEBES e FEELDERS, 2003).

## 3.2 ITENS FREQÜENTES

Uma regra de associação ($X \Rightarrow I$) é composta por um antecedente ($X$) e um conseqüente ($I$), ambos denominados *itemset*. Um conjunto de itens é referido como um *itemset*[1]. Um *itemset* que contém *k* itens é um *k-itemset*, por exemplo, o conjunto {*A*, *B*} é um *2-itemset*. A freqüência de ocorrência de um *itemset* é o número de transações que contém o *itemset*. O número de transações requeridas para que um *itemset* satisfaça o suporte mínimo é, portanto, referido como suporte mínimo. Um *itemset*[2] será chamado de **item freqüente** nesta dissertação. O conjunto de *k-itemset* freqüente é denotado por $L_k$[3] (HAN e KAMBER, 2001).

[1] Na literatura da pesquisa de mineração de dados, "*itemset*" é mais comumente usado do que "*item set*".

[2] Nos trabalhos passados, *itemsets* que satisfazem o suporte mínimo eram referenciados como *large*. Esse termo, entretanto, causa confusão visto que ele tem uma conotação ao número de itens de um *itemset* ao invés da freqüência de ocorrências do conjunto. Dessa forma, mais recentemente, utilizamos o termo *freqüente*.

[3] Embora o termo *freqüente* seja melhor que *large*, por razões históricas, um conjunto de *k-itemset* freqüente continua sendo denotado como $L_k$. (***L* - *L**arge***)

## 3.3 MEDIDAS DE INTERESSE

Um dos principais problemas no campo da descoberta do conhecimento é o desenvolvimento de boas medidas de interesse para a descoberta de padrões de dados (SILBERSCHATZ e TUZHILIN, 1995). Tais medidas de interesse são classificadas em medidas objetivas e subjetivas. As medidas objetivas são aquelas que dependem apenas da estrutura do padrão e do dado latente usados no processo de descoberta, e podem lidar com técnicas que não requerem conhecimento da aplicação ou domínio. As medidas subjetivas são aquelas que dependem, também, da classe de usuários que examinam o padrão e medem o interesse subjetivo dos padrões para o usuário (LIU et al., 1999), tais como: *unexpectedness* (SILBERSCHATZ e TUZHILIN, 1995) e *actionability* (PIATETSKY-SHAPIRO e MATHEUS, 1994).

Alguns dos principais fatores que contribuem para o interesse de um padrão descoberto são: *rule templates* (KLEMETTINEN et al., 1994), *metarule-guided* (KAMBER et al., 1997), *neighborhood-based interestingness* (DONG e LI, 1998), *lift* (REINSCHMIDT et al., 1999)(*strenght* (DHAR e TUZHILIN, 1993) *e interest* (BRIN et al., 1997)), *conviction* (BRIN et al., 1997), *collective strength* (AGGARWAL e YU, 1998), *gain* (FUKUDA et al., 1996), *entropy gain* (MORISHITA, 1998), *coverage* (QUINLAN, 1992), *leverage* (PIATETSKY-SHAPIRO, 1991) e *any-confidence, all-confidence* e *bond* (OMIECINSKI, 2003). Essas medidas podem ser utilizadas para a decisão de quais regras devem ser mantidas e quais devem ser descartadas (LIU et al., 1999). O processo de mineração exibe ao usuário apenas aquelas mais interessantes. Dessa forma, reduz-se o tempo de análise das regras pelos usuários.

A seguir, são descritas algumas das medidas de interesse mencionadas:

- *actionability*: é baseada na utilidade da regra para o usuário. É uma importante medida de interesse subjetiva, porque os usuários estão mais interessados no conhecimento que permite que eles melhorem seu trabalho, tomando algumas ações em resposta ao novo conhecimento descoberto;
- *unexpectedness*: se um novo padrão é descoberto, então é surpreendente para o usuário, o que significa que o conhecimento contradiz a convicção dos usuários. Dessa forma, a medida *unexpectedness* está intimamente relacionada às convicções, crenças ou impressões gerais. A crença ou convicção pode ser classificada em dois tipos: *hard belief*, restrições que não podem ser alteradas com novas evidências, e *soft belief*, crenças que o usuário gostaria que mudassem com a nova evidência;
- *rule template*: classes de regras interessantes e não interessantes podem ser especificadas com *templates*. *Templates* descrevem um conjunto de regras especificando quais atributos ocorrem no antecedente e no conseqüente. Com o uso de *templates*, o usuário pode especificar explicitamente tanto o que é interessante quanto o que não é. Para ser interessante, a regra tem que se igualar ao *template* inclusivo (*inclusive template*). Se uma regra, entretanto, se igualar a um *template* restritivo (*restrictive template*), é considerada desinteressante;
- *neighborhood-based unexpectedness*: foi proposta como uma forma de identificar regras interessantes. Nessa abordagem, o interesse de uma regra depende não apenas do seu próprio suporte ou confiança, mas também do suporte e confiança das regras que estão na sua vizinhança;
- *collective strength*: foi projetada para mostrar como um valor de um atributo afeta o valor de outro. O item freqüente *I* é denotado como *stronglycollective* no nível *K*, se ele satisfaz as seguintes propriedades: a *collective strength C(I)* de um item do conjunto de itens

freqüentes *I* é pelo menos *K*, e a *collective strength C(J)* de todo subconjunto *J* de *I* é pelo menos *K*. Um *itemset I* está em violação de uma transação, se alguns itens estão presentes em uma transação e os outros não. O índice de violação de um *itemset I* é denotado por *v*(*I*) e é a fração de violações do *itemset I* sobre todas as transações. O *collective strength* de um *itemset I* é denotado por *C*(*I*) e é definido conforme a seguir:

$$C(I) = \frac{1 - v(I)}{1 - E[v(I)]} \cdot \frac{E[v(I)]}{v(I)}$$

- *lift* (*strenght* ou *interest*): é definida como o fator real de confiança dividido pela confiança esperada. O interesse de *A, B* é definido como *P*(*A,B*)/*P*(*A*)*P*(*B*). É uma medida do desvio da previsão para a expectativa de uma regra. Se, por exemplo, todo mundo comprar chiclete na caixa registradora, haverá muitas regras que associam todos os tipos de produto com chiclete, e, portanto, um suporte e confiança altos podem não ter significância. Nesse exemplo, *lift* diz o quão alto é o nível de confiança (comparado ao percentual total de transações) quando o segundo item está presente. Um *lift* alto significa que a conexão entre os itens é forte. O *lift* pode ser usado para qualificar a confiança como relevante ou não;
- *conviction*: tem sempre o valor 1 (um) quando os itens freqüentes não se relacionam, ou seja, são completamente independentes, e ∞ quando os itens sempre aparecem juntos. Enquanto outras medidas simétricas, como *lift*, medem a dependência entre os itens, ou seja, *lift*(A⇒B) = *lift*(B⇒A), a *conviction* é direcional e extremamente afetada pela direção da seta, ou seja, *conviction*(A⇒B) ≠ *conviction*(B⇒A). Quanto mais alta a *conviction*, mais freqüente o antecedente ocorre com o conseqüente. Se for igual a 1 (um) indica independência. *Conviction* é definida como $P(A)P(\neg B)/P(A,\neg B)$ ; e

- *any-confidence, all-confidence* e *bond*: com a medida *any-confidence*, uma associação é considerada interessante se nenhuma regra que pode ser derivada daquela associação tem uma confiança maior ou igual ao valor do *any-confidence* mínimo. Nos algoritmos de mineração de regras de associação, é o mesmo que dizer que queremos todas as regras que têm a confiança maior ou igual à confiança mínima, sem levar em conta nenhum critério de suporte. Com a medida *all-confidence*, uma associação é considerada interessante se todas as regras que podem ser produzidas daquela associação tiverem uma confiança maior ou igual ao valor do *all-confidence* mínimo. A medida *bond* é similar ao suporte, a diferença está no fato de que o suporte se refere ao conjunto inteiro do dado e o *bond* se refere ao subconjunto do dado. A idéia dessa medida é encontrar todos os conjuntos de itens que são freqüentes em um conjunto de dados definidos de acordo com uma característica do dado.

  Considere a definição de um conjunto *m* itens *I* como $\{i_1,i_2,...,i_m\}$ e o conjunto de transações em *I* como *D*. Cada transação contém um conjunto de itens que são subconjuntos de *I*. Considere, também, a notação utilizada para definir o suporte de um conjunto de itens $\mathcal{L}$ :

$$\frac{|\{d \mid d \in \mathcal{D} \wedge \mathcal{L} \subset d\}|}{|\mathcal{D}|}.$$

A definição de *any-confidence* é:

$$\frac{|\{d \mid d \in \mathcal{D} \wedge \mathcal{L} \subset d\}|}{MIN\{i \mid \forall l(l \in \mathcal{P}(\mathcal{L}) \wedge l \neq \emptyset \wedge l \neq \mathcal{L} \wedge i = |\{d \mid d \in \mathcal{D} \wedge l \subset d\}|)\}}.$$

A definição de *all-confidence* é:

$$\frac{|\{d \mid d \in \mathcal{D} \wedge \mathcal{L} \subset d\}|}{MAX\{i \mid \forall l(l \in \mathcal{P}(\mathcal{L}) \wedge l \neq \emptyset \wedge l \neq \mathcal{L} \wedge i = |\{d \mid d \in \mathcal{D} \wedge l \subset d\}|)\}}.$$

A definição de *bond* é:

$$\frac{|\{d \mid d \in \mathcal{D} \wedge \mathcal{L} \subset d\}|}{|\{d \mid d \in \mathcal{D} \wedge \exists l(l \in \mathcal{P}(\mathcal{L}) \wedge l \neq \emptyset \wedge l \subset d)\}|}.$$

## 3.4 ALGORITMOS DE MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO

Os algoritmos do tipo *Apriori* para descoberta de itens freqüentes percorrem a base de dados diversas vezes (AGRAWAL e SRIKANT, 1994). No primeiro passo do algoritmo, o suporte de cada conjunto de itens é contado individualmente e determina dentre eles quais são os freqüentes, isto é, quais possuem o suporte mínimo especificado. Neste passo, a quantidade de itens do conjunto de itens é um ($k$=1). Utilizando os itens freqüentes do passo anterior e algum critério específico do algoritmo, em cada passo subseqüente são gerados os novos potenciais itens freqüentes, chamados de conjuntos de itens candidatos. Neste passo, um elemento é adicionado ao conjunto de itens ($k$+1). Para cada um dos conjuntos de itens candidatos, o suporte é calculado ao percorrer a base de dados novamente. No final de cada passo, são determinados quais itens candidatos são realmente freqüentes. Estes, por sua vez, servirão para a geração dos novos candidatos. Esse processo continua até que não exista mais nenhum conjunto de itens freqüentes (AGRAWAL e SRIKANT, 1994).

Vários algoritmos para descoberta de itens freqüentes foram propostos, tais como: AIS (AGRAWAL et al., 1993), SETM (HOUTSMA e SWAMI, 1993), *Basic* (KLEMETTINEN et al., 1994), *Apriori*, *AprioriTid*, *AprioriHybrid* (AGRAWAL e SRIKANT, 1994), *FP-Grouwth* (HAN et al., 2000), *Dynamic Itemset Counting* (DIC) (BRIN et al., 1997), *Carma* (HIDBER, 1999), FPMax (GRAHNE e ZHU, 2003), *Direct Count and Intersect Closed* (DCI *Closed*) (LUCCHESE et al., 2004), *GenMAX* (GOUDA e ZAKI, 2001), *Tree Projection* (AGRAWAL

et al., 2001), MAFIA (BURDICK et al., 2001) e *LPMiner* (SENO e KARYPIS, 2001). Dentre esses algoritmos destacamos o *Apriori* devido a sua relevância para esta pesquisa, tendo em vista o fato dele possuir ciclos, como será exposto no Capítulo 5. Esse algoritmo será detalhado a seguir.

## 3.5 ALGORITMO *APRIORI*

O algoritmo *Apriori* é um algoritmo básico, definido por Agrawal et. al. (1994), utilizado para encontrar itens freqüentes em um banco de dados. Seu nome é baseado no fato do algoritmo usar conhecimento prévio (*prior knowledge*) das propriedades do item freqüente. Emprega uma abordagem iterativa conhecida como busca *level-wise*, onde *k-itemsets* são usados para explorar *(k+1)-itemsets*.

### 3.5.1 CONCEITOS BÁSICOS

Seja $\tau = \{i_1, i_2, ..., i_m\}$ um conjunto de itens. Seja $D$, um conjunto de transações de um banco de dados onde cada transação $T$ é um conjunto de itens tal que $T \subseteq \tau$. Cada transação é associada a um identificador, chamado *TID*. Seja $A$ um conjunto de itens. Uma transação $T$ é contém $A$ se e apenas se $A \subseteq T$. Uma regra de associação é uma implicação na forma de $A \Rightarrow B$, onde $A \subset \tau$, $B \subset \tau$, e $A \cap B = \varnothing$. A regra $A \Rightarrow B$ faz parte do conjunto de transação $D$ com suporte $s$, onde $s$ é o percentual de transações em $D$ que contém $A \cup B$ (isto é, ambos $A$ e $B$). Isso é tido como probabilidade $P(A \cup B)$. A regra $A \Rightarrow B$ tem confiança $c$ no conjunto de transação $D$, se $c$ é o percentual de transações em $D$ contendo $A$ que também contém $B$. Isso é tido como a probabilidade condicional $P(B|A)$ (HAN e KAMBER, 2001). Isto é: suporte($A \Rightarrow B$) = $P(A \cup B)$ e confiança($A \Rightarrow B$) = $P(B|A)$. Nesta dissertação, os itens de um conjunto de itens contêm o nome de um campo do banco de dados e o seu respectivo valor, conforme o modelo de dados apresentado no item 6.2 e o exemplo de aplicação do capítulo 7.

Regras que satisfazem tanto um suporte mínimo quanto uma confiança mínima são chamadas regras fortes (*strong rules*). Por convenção, os valores de suporte e confiança são escritos como valores percentuais.

O algoritmo *Apriori* é executado em dois passos:

1. Encontrar todos os itens freqüentes: por definição, cada um desses itens freqüentes ocorre tão freqüentemente quanto o suporte mínimo pré-determinado; e
2. Gerar regras de associação fortes a partir dos itens freqüentes: por definição, essas regras devem satisfazer à confiança mínima.[4]

A propriedade *Apriori* (*Apriori property*) é usada para reduzir o espaço de busca pelos itens freqüentes, baseada na seguinte observação (HAN e KAMBER, 2001):

- Se um *itemset I* não satisfaz `min-sup`, então *I* não é freqüente: P(*I*) < `min-sup`; e
- Se o item *A* for adicionado ao *itemset I*, então o *itemset* resultante ($I \cup A$) não pode ocorrer mais freqüente que *I* : P($I \cup A$) < `min-sup`.

Essa propriedade pertence a uma categoria especial de propriedades chamada *anti-monotone* que afirma: se um conjunto não pode passar em um teste, todos os seus superconjuntos irão falhar no mesmo teste.

A propriedade *Apriori* é usada no algoritmo no processo para encontrar $L_k$ a partir de $L_{k-1}$, por meio dos seguintes passos: *join* e *prune*.

- *join*: para encontrar $l_k$, um conjunto de *k-itemsets* é gerado aplicando um *join* de $L_{k-1}$ com ele mesmo. Esse conjunto candidato é denotado como $C_k$. $l_1$ e $l_2$ são *itemsets* em $L_{k-1}$. $l_{i[j]}$ refere-se ao *j-ésimo* item em $l_i$, ou seja, $l_{1[k-2]}$ refere-se ao segundo item de trás para frente em $l_i$; e

[4] Medidas de interesse adicionais podem ser aplicadas.

- *prune*: $C_k$ é um super conjunto de $L_k$. Qualquer *(k-1)-itemset* que não é freqüente não pode ser um subconjunto de um *k-itemset* freqüente.

### 3.5.2 DESCRIÇÃO DO ALGORITMO

O algoritmo *Apriori* possui ciclos evolutivos claros e delimitados no processo de mineração de itens freqüentes, característica fundamental para a implementação de um processo de mineração interativo, que é o propósito desta dissertação. A Figura 2 mostra o algoritmo *Apriori*. O primeiro passo do algoritmo, simplesmente, conta as ocorrências para determinar a freqüência de *1-itemset* e gerar $L_1$. $L_1$ é o conjunto de *1-itemset* freqüente que possui suporte maior que o suporte mínimo. Cada membro desse conjunto tem dois campos: item freqüente e suporte. O passo subseqüente, chamado passo *k*, consiste em duas fases: na primeira, os itens freqüentes de $L_{k-1}$ encontrados no (*k-1*)-ésimo passo são usados para gerar a tabela de itens candidatos $C_k$, usada na função `apriori-gen`, descrita na próxima seção; e na segunda, o banco de dados é varrido e o suporte dos candidatos em $C_k$ é calculado. Para um cálculo rápido, é necessária a determinação dos candidatos em $C_k$ que estão contidos em uma dada transação *t*.

```
L1 = {large 1-itemsets);
for (k = 2; Lk-1 <> 0; k++) do begin
      Ck = apriori-gen(Lk-1); // Novas candidatas
      forall transactions t ∈ D do begin
         Ct = subset(Ct,t); // Candidatas contidas em t
         forall candidates c ∈ Ct do
            c.count ++;
      end
      Lk = {c E Ck | c.count ≥ minsup}
end
Answer = ∪k Lk;
```

**Figura 2: Algoritmo *Apriori***

### 3.5.3 TABELA CANDIDATA

A tabela candidata é gerada pela função `apriori-gen` do algoritmo *Apriori*. Essa função recebe o argumento $L_{k-1}$, isto é, o conjunto de todos os *(k-1)-itemsets* freqüentes, e retorna um superconjunto do conjunto de todos os *k-itemsets* freqüentes. A função `apriori-gen` possui dois passos: *Join* e *Prune*. No primeiro, $L_{k-1}$ é cruzada com $L_{k-1}$; no segundo, são apagados todos os *itemsets* $C \in C_k$ tais que alguns *(k-1)-subset* de $C$ não estão contidos em $L_{k-1}$.

- *join*: $C_k$ é gerada pela junção de $L_{k-1}$ com ela mesma.

```
insert into Ck
    select p.item1, p.item2, …, p.itemk-1, q.itemk-1
      from Lk-1 p, Lk-1 q
      where p.item1=q.item1, …, p.itemk-2=q.itemk-2, p.itemk-1 < q.itemk-1
```

- *prune* (corta, poda): qualquer *(k-1)-itemset* que não é freqüente, não pode ser um subconjunto de um *k-itemset* freqüente.

```
forall itemsets c in Ck do
      forall (k-1)-subsets s of c do
            if (s is not in Lk-1) then delete c from Ck
```

### 3.5.4 GERAÇÃO DE REGRAS DE ASSOCIAÇÃO A PARTIR DE ITENS FREQÜENTES

Uma vez que os itens freqüentes em um banco de dados foram encontrados, o passo seguinte é gerar as regras de associação que satisfazem tanto o suporte mínimo quanto a confiança mínima. Isso pode ser feito usando a equação de confiança, onde a probabilidade condicional é expressa em termos do valor do suporte do item freqüente (HAN e KAMBER, 2001):

- confiança *(A ⇒ B) = P(B | A)* = suporte *(A ∪ B) /* suporte *(A),* onde suporte (*A ∪ B*) é o número de transações contendo os *itemset A ∪ B*, e suporte (*A*) é o número de transações contendo o *itemset* A.

Baseado nessa equação, regras de associação podem ser geradas, como a seguir:

- Para cada *itemset* freqüente *l*, gerar todos os subconjuntos não-vazios de *l*; e
- Para cada conjunto não-vazio *s* de *l*, gerar a regra "*s ⇒ (l-s)*" se suporte(*l*)/suporte(*s*) ≥ `min-conf`, onde `min-conf` é a confiança mínima.

# 4 ABORDAGENS PARA EXTRAÇÃO E MANIPULAÇÃO DE REGRAS DE ASSOCIAÇÃO

Vários métodos de descoberta de padrões propostos na literatura técnica têm o inconveniente da descoberta de padrões óbvios e irrelevantes (PADMANABHAN e TUZHILIN, 1999). Em (SIEBES e FEELDERS, 2003) são propostas duas direções para gerenciar o excesso de resultados: o pré e o pós-processamento das regras. O pré-processamento gera menos regras, como as medidas de interesse definidas no item 3.3 e a abordagem para o Direcionamento e Redução do Número de Regras do item 4.2 a seguir. O pós-processamento filtra ou ordena as regras de forma que aquelas mais interessantes sejam selecionadas, como o OLAP *Mining*, o DWARF e as Técnicas de Visualização. Essas abordagens foram propostas para amenizar o problema da geração do excesso de regras de associação e serão apresentadas a seguir.

## 4.1 OLAP *MINING*

O conceito de OLAP *Mining* é definido em (HAN, 1997) como o mecanismo que integra o processamento analítico *on-line* com a mineração de dados de tal forma que as operações de mineração possam ser efetuadas sobre diferentes porções de um banco de dados, ou de um *data warehouse*, em diferentes níveis de abstração, através de simples requisições do usuário. As funções de *cubing* (multidimensionalizar) e *mining* (minerar) podem ser intercaladas e integradas para fazer da mineração de dados um processo interessante e de alta interatividade. Han (1997) propõe que as seguintes funcionalidades estejam presentes em um OLAP *Mining*:

- *cubing then mining*: com a disponibilidade de cubos de dados e operações sobre cubo, a mineração pode ser executada em qualquer camada e em qualquer porção de um cubo de

dados. Desse modo, as operações sobre cubo podem selecionar qual parte do cubo e quais níveis de abstração (granularidade) serão utilizados antes do início do processo de mineração;

- *mining then cubing*: a mineração pode ser, primeiramente, executada em um cubo de dados, e, posteriormente, resultados particulares da mineração podem ser analisados pelas operações de cubo;
- *cubing while mining*: um modo flexível de integrar operações de *mining* e *cubing* é a execução de operações de mineração similares em múltiplas granularidades. Dessa forma, as mesmas operações de mineração podem ser executadas em diferentes partes do cubo ou em diferentes níveis de abstração;
- *backtracking*: usado para facilitar a mineração interativa, permite que o processo de mineração volte um ou dois passos ou volte a um marco definido e, então, explore caminhos de mineração alternativos; e
- *comparative mining*: permite a mineração de dados comparativos, isto é, a comparação entre diferentes processos de mineração de dados.

Uma arquitetura integrada de *On-Line Analytical Mining* (OLAM) e OLAP é definida em (HAN e KAMBER, 2001), onde a mineração analítica é executada em cubos de dados de maneira similar a um processamento analítico (Figura 3). Han e Kamber (2001) acreditam que mineração de dados deve ser um processo centrado no humano (*human-centered*). Assim, ao invés de uma solicitação ao sistema de mineração de dados para gerar padrões e conhecimento, automaticamente, o usuário precisa interagir freqüentemente com o sistema para realizar análise e exploração do dado.

Han (1997) afirma que em *data warehouses* contendo um grande volume de dados é crucial o fornecimento de flexibilidade na mineração de dados para que o usuário possa manipular um cubo de dados, selecionar o espaço de mineração e os níveis de abstração desejados, e testar diferentes módulos de mineração e algoritmos alternativos.

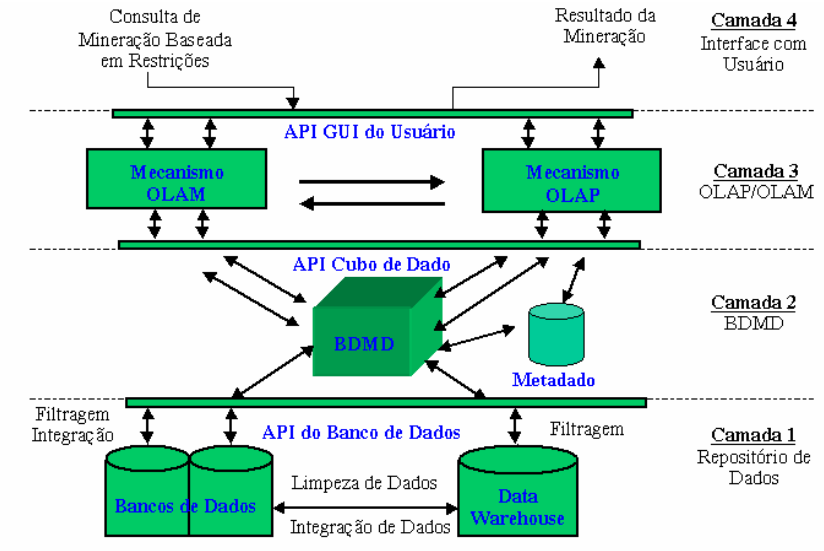



**Figura 3: Uma arquitetura integrada de OLAM e OLAP**
**Fonte: (HAN e KAMBER, 2001)**

## 4.2 DIRECIONAMENTO E REDUÇÃO DO NÚMERO DE REGRAS

Nessa abordagem foi implementado o Serviço de Geração de Regras de Associação baseado no algoritmo *Apriori* (LAVÔR, 2003). Esse serviço utiliza diversos parâmetros para reduzir o número de regras de associação geradas. O objetivo da redução do número de regras geradas é possibilitar ao usuário a manipulação de um volume menor de informações ao final da execução do algoritmo. Segundo Lavôr (2003), os seguintes parâmetros estão aptos a direcionar o resultado da mineração: número máximo de itens de um conjunto de itens

freqüentes; quais itens deverão ser utilizados no antecedente das regras e quais itens deverão ser utilizados no conseqüente das regras; e o número máximo de regras a serem geradas.

Além desses parâmetros, as combinações de itens que são triviais para determinado problema podem ser enviadas ao serviço implementado para serem descartadas. As combinações podem ser enviadas, mesmo que não sejam triviais, desde que o usuário não tenha interesse em regras que possuam determinados itens juntos. As combinações triviais passadas como parâmetro são descartadas pelo serviço na etapa de geração de itens candidatos.

## 4.3 DWARF

O *Data Warehouse Association Rule Framework* (DWARF), proposto em (MONTEIRO et al., 2003), define um *data warehouse* de regras de associação através da especificação de estruturas de dados, operações e modelos capazes de prover um ambiente para análise de regras de associação. Essa abordagem se beneficia naturalmente de muitas características de *data warehouse* e corresponde a uma estrutura interessante para propor novos modelos de descoberta de conhecimento.

Nesse trabalho, um modelo de descoberta de conhecimento corresponde à inferência de possíveis razões para a mudança de regras ao longo do tempo. Essa é uma informação preciosa para entender mudanças no comportamento de um negócio.

Um *data warehouse* de regras de associação deve possuir as seguintes características (MONTEIRO et al., 2003):

- Permitir a visualização de diferentes valores agregados referentes a um conjunto de regras de associação, desde uma simples contagem até métricas mais elaboradas ou medidas de

interesse. A visualização do conjunto de regras também deve ser possível, seja através de uma simples listagem ou de técnicas mais elaboradas de visualização;

- Permitir a associação de medidas de interesse a dimensões do *data warehouse*, tais como suporte e confiança. Como exemplo, deve ser possível visualizar o total de regras existentes por faixa de suporte ou restringir as regras consideradas a uma faixa qualquer;
- Possuir uma dimensão Tempo seguindo a abordagem tradicional de *data warehouse*. A aplicação de uma restrição qualquer nessa dimensão deve gerar o efeito de considerar apenas regras válidas no período de tempo sendo analisado;
- Possuir duas dimensões Itens, uma associada ao antecedente da regra e outra ao conseqüente. Essas dimensões podem ter hierarquias permitindo diversas organizações e agrupamentos de itens;
- Permitir a execução de operações OLAP tradicionais, tais como *drill-down*, *roll-up*, *pivot*, *slice*, *dice*, etc, sobre o cubo de regras sendo analisado; e
- Possuir outras dimensões como, por exemplo, uma dimensão espacial, que permita a restrição de regras válidas em uma determinada região.

Os dados necessários às operações de um *data warehouse* de regras de associação deverão estar disponíveis em alguma estrutura carregada previamente, a exemplo de um *data warehouse* convencional.

## 4.4 TÉCNICAS DE VISUALIZAÇÃO

Em (BRUZZESE e BUONO, 2004) são apresentadas técnicas baseadas em gráficos e coordenadas paralelas para visualização dos resultados dos algoritmos de mineração de regras de associação. É importante ressaltar que as técnicas de visualização não reduzem o número de regras, mas facilitam o seu entendimento, ao permitir a escolha das regras mais pertinentes.

A demanda por ferramentas de análise visuais e interativas é latente no contexto de regras de associação onde, freqüentemente, os usuários têm que analisar centenas de regras com o intuito de capturar conhecimento valioso.

Em (BRUZZESE e BUONO, 2004) é apresentada uma estratégia visual para tratar do inconveniente do excesso de regras. Essa estratégia explora uma técnica baseada em gráfico e em coordenadas paralelas para visualizar os resultados dos algoritmos de mineração de regras de associação. Na Figura 4, um subconjunto de 774 regras caracterizadas pelo mesmo conseqüente é exibido. O usuário pode decidir qual subconjunto visualizar pela seleção de um item conseqüente de um *listbox*; o sistema automaticamente mostra, no espaço das coordenadas paralelas, todas as regras com aquele conseqüente e adiciona duas dimensões suplementares para medidas de suporte e confiança.

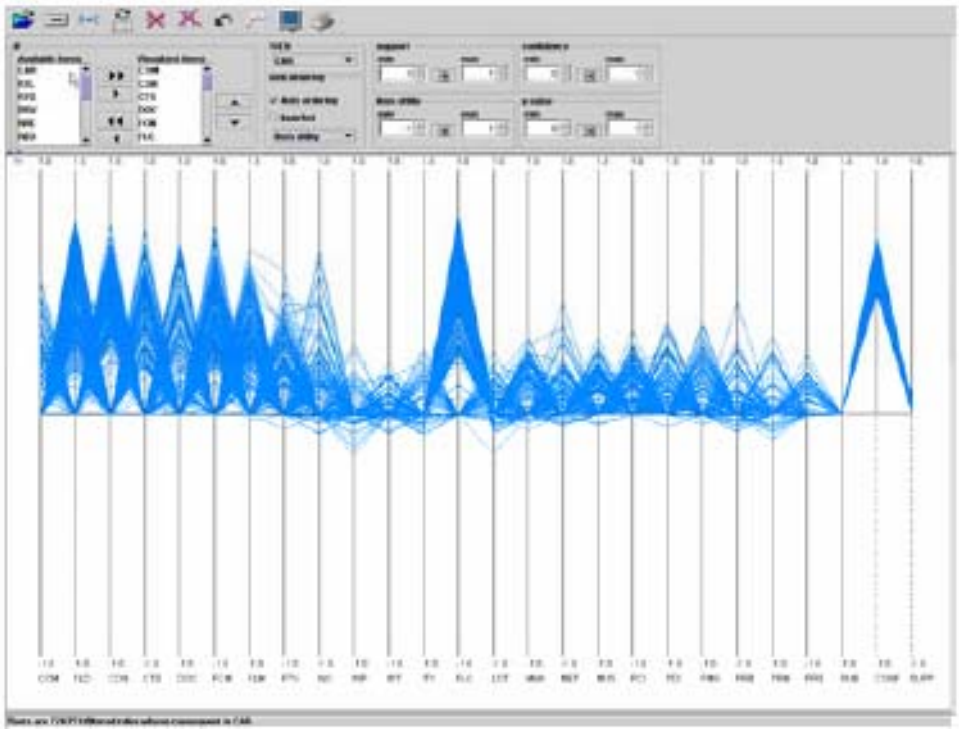


**Figura 4: Visualização de Regras de Associação Utilizando Coordenadas Paralelas**
**Fonte: (BRUZZESE e BUONO, 2004)**

Outros exemplos de visualização de regras de associação estão ilustrados nas Figuras 5 e 6 a seguir.

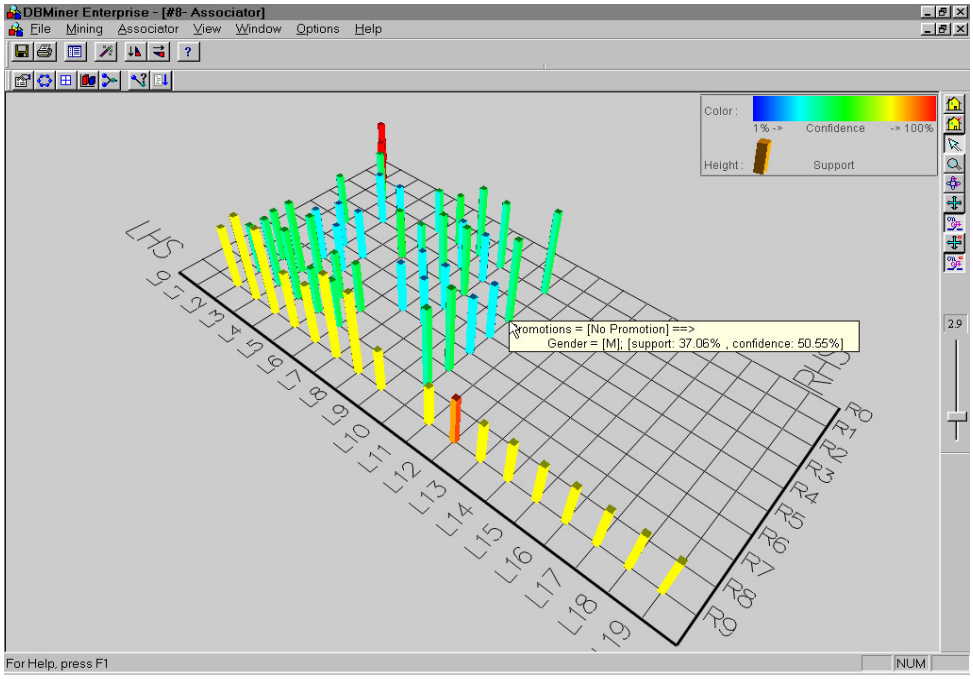



**Figura 5: Visualização de Regras de Associação Utilizando Gráfico Plano**
**Fonte: (HAN e KAMBER, 2001)**

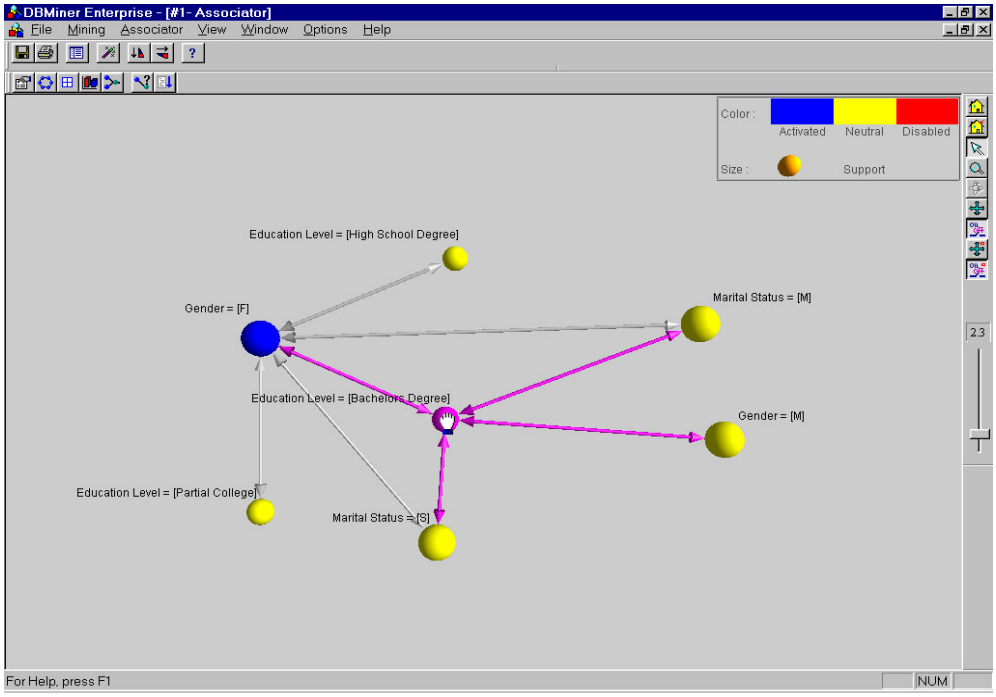



**Figura 6: Visualização de Regras de Associação Utilizando Grafo de Regras**
**Fonte: (HAN e KAMBER, 2001)**

## 4.5 DISCUSSÃO DAS ABORDAGENS APRESENTADAS

A Tabela 1 apresenta uma comparação entre a abordagem proposta e algumas abordagens encontradas na literatura (Capítulo 4), observando as características de padrão de dado interessante: fácil de entender por humanos, potencialmente útil e novo (HAN e KAMBER, 2001). Observa, também, as características do processo: permitir a interferência do usuário com vistas à obtenção das regras mais interessantes e fornecer funcionalidades analíticas próprias de um *data mart*; e a modelagem de dados multidimensional para regras de associação.

**Tabela 1: Abordagens x Características**

| Abordagens | Características de Padrão Interessante | | | Características do Processo | | Modelo Multi-dimensional |
|---|---|---|---|---|---|---|
| | Fácil de entender | Potencial-mente útil | Potencial-mente novo | Interferência do usuário | Funcionalida-des analíticas | |
| OLAP *Mining* | ✓ | ✓ | ✓ | × | ✓ | × |
| Medidas de Interesse | × | ✓ | ✓ | × | × | × |
| Direc. e redução | × | ✓ | ✓ | × | × | × |
| Téc. de Visualização | ✓ | NA | NA | NA | NA | NA |
| DWARF | ✓ | ✓ | ✓ | × | ✓ | × |
| Ambiente Analítico Proposto | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

NA - Não se aplica; ✓ - Sim; × - Não

Observa-se que (i) nenhuma das abordagens apresentadas permite que o usuário interfira no processo de mineração e nenhuma define um modelo multidimensional das regras de associação; (ii) apenas a OLAP *Mining* e o DWARF incluem funcionalidades analíticas para manipulação das regras; e (iii) a facilidade de entendimento das regras pelo usuário é uma preocupação apenas da OLAP *Mining*, das Técnicas de Visualização e do DWARF.

A abordagem analítica proposta visa suprir pontos onde as outras abordagens se apresentaram incompletas, reunindo todas as características analisadas. O ambiente e o processo analítico para exploração de regras de associação proposto será apresentado no próximo capítulo.

# 5 UMA ARQUITETURA DE EXPLORAÇÃO DE REGRAS DE ASSOCIAÇÃO

A abordagem alternativa proposta neste trabalho para o tratamento do excesso de regras de associação geradas pela tarefa de mineração é a definição de uma arquitetura de exploração de regras em um ambiente analítico, como inicialmente visto em (HAN, 1997), onde as funcionalidades OLAP se integram ao processo de mineração para que regras de associação mais pertinentes possam ser geradas e, posteriormente, analisadas e manipuladas pelo usuário. A integração das funcionalidades OLAP com o processo de mineração é feita por meio da análise dos resultados parciais gerados a cada ciclo do algoritmo *Apriori* (seção 3.5), permitindo que novas medidas de interesse e parâmetros de entrada sejam definidos e, assim, que a mineração seja direcionada.

No ambiente analítico proposto, o repositório de dados original, o resultado final da mineração e os resultados parciais de cada um dos seus ciclos são disponibilizados em uma ferramenta OLAP. Essa ferramenta permite a análise do processo de exploração sob diferentes perspectivas. Dessa forma, o usuário pode visualizar o caminho que o algoritmo está percorrendo e redirecionar a mineração para as informações sobre as quais deseja novos conhecimentos. Pode, também, selecionar as regras mais interessantes e confrontá-las com os dados originais.

Um modelo de dados multidimensional foi projetado para contemplar a memória do processo de exploração, que contém os itens freqüentes encontrados, as regras selecionadas e descartadas para o próximo ciclo, os parâmetros utilizados para interferir na seleção das regras que permanecerão na mineração. Todas essas informações e o resultado da mineração são armazenados em um cubo, onde cada atributo dos itens freqüentes das regras de associação encontradas é referenciado como uma dimensão do modelo de dados.

Dessa forma, a arquitetura proposta apóia o analista de negócios no processo de tomada de decisão oferecendo um ambiente fácil de usar e entender. Entretanto, mesmo com um ambiente amigável, é necessário que o usuário tenha conhecimento técnico e específico do processo de exploração de regras de associação e habilidades para operação de uma ferramenta OLAP.

## 5.1 ELEMENTOS DA ARQUITETURA

A arquitetura de exploração de regras de associação é composta pelos seguintes elementos (Figura 7): **repositório de dados**, contendo os dados originais; **porção do repositório de dados**, contendo a seleção dos dados que será minerada; s**erviço de mineração de regras de associação,** que implementa o algoritmo *Apriori*; **medidas de interesse e parâmetros de entrada,** definidas pelo usuário para interferir no processo de mineração; **ferramenta OLAP**, que permite a realização de análises sobre as regras geradas e o repositório de dados; **regras de associação e informações de controle**, resultantes dos ciclos do algoritmo de mineração, que poderão ser, também, analisadas pelo usuário.

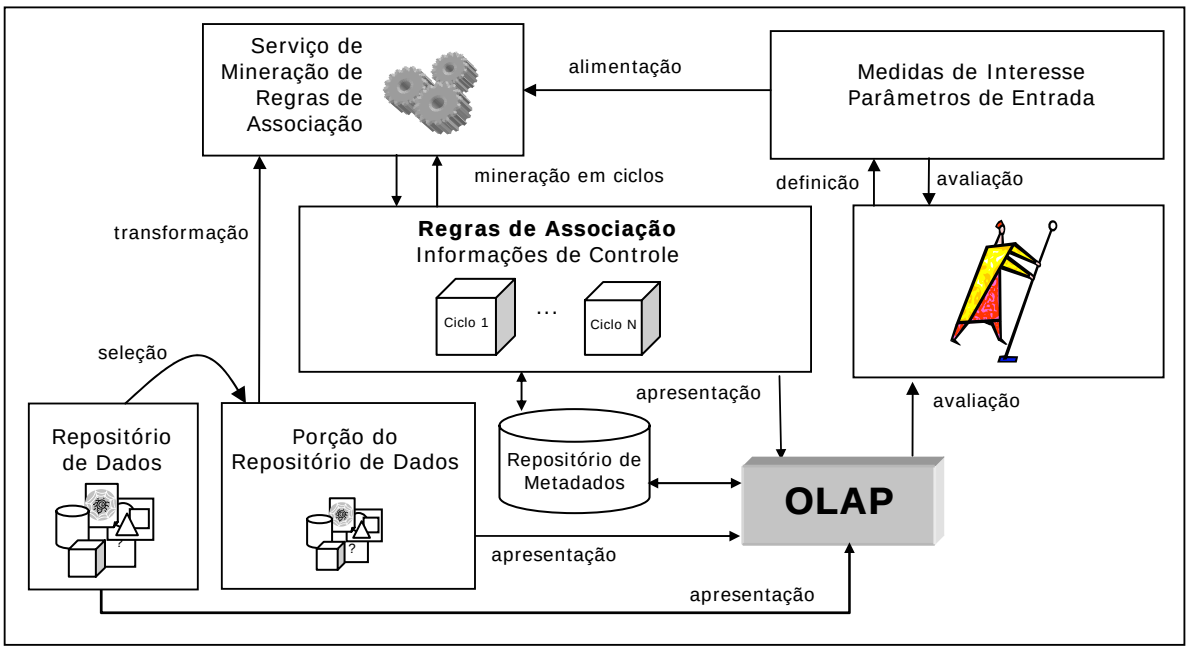



**Figura 7: Arquitetura do Ambiente Analítico**

Os elementos da arquitetura fazem parte de um ambiente integrado, onde a conexão entre eles e entre os repositórios de dados é transparente para o usuário. Esses elementos serão descritos nas seções a seguir.

### 5.1.1 REPOSITÓRIO DE DADOS

Padrões de dados podem ser minerados a partir de diferentes tipos de banco de dados, como: bancos de dados relacionais, *data warehouses*, bancos de dados transacionais, relacional-objeto e orientado-a-objeto. Padrões de dados interessantes podem, também, ser extraídos de outros tipos de repositórios de dados, incluindo, bancos de dados espaciais, temporais, textuais, multimídia, bancos de dados legados e dados da *World Wide Web*. O repositório de dados pode ser usado para apoiar o usuário durante o processo de exploração, por meio do seu confronto com a memória da mineração.

O repositório de dados é oriundo de um ambiente transacional ou de um ambiente analítico, com os dados consistentes e sem ruídos, onde múltiplas fontes de dados foram combinadas. Na arquitetura proposta, uma porção do repositório de dados é selecionada a partir do repositório de dados original, com a utilização de uma ferramenta própria do repositório. Essa porção contém apenas os dados que o usuário considera relevantes para o processo de descoberta de padrões e pode ser redefinida a cada ciclo da mineração. No exemplo de aplicação (Capítulo 7), o repositório de dados utilizado foi o banco de dados do vestibular da UFRJ. Alguns campos desse banco foram selecionados e processados pelo Serviço de Mineração de Regras de Associação. A seleção dos campos foi alterada durante todo o processo de exploração de acordo com o assunto a ser minerado, por exemplo, os campos relativos ao aspecto sócio-cultural do candidato foram minerados separadamente dos campos relativos ao aproveitamento dos candidatos no vestibular.

### 5.1.2 SERVIÇO DE MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO

O Serviço de Mineração de Regras de Associação da arquitetura proposta utiliza medidas de interesse e parâmetros de entrada para procurar por padrões na porção do repositório de dados definida pelo usuário. Esse serviço implementa o algoritmo *Apriori*, escolhido pela sua simplicidade e característica de possuir ciclos. A cada ciclo, as regras candidatas são geradas e passam por uma crítica. Apenas as regras que estão de acordo com as medidas de interesse e os parâmetros de entrada inicialmente definidos passam para o ciclo seguinte, sendo as demais descartadas. Nesta dissertação, chamamos de regras candidatas àquelas que possuem itens freqüentes (atendem ao suporte mínimo especificado), mas não atendem às outras medidas de interesse ou parâmetros de entrada. Essas regras estarão disponíveis para o usuário avaliá-las. A partir dessa avaliação o usuário pode decidir pela sua permanência no processamento. Desta forma, ele deverá alterar as medidas de interesse ou os parâmetros de entrada que descartaram a regra para que ela conste no resultado da mineração.

Parte do trabalho desta dissertação foi adaptar a implementação do Serviço de Geração de Regras de Associação de Lavôr (2003), baseado no algoritmo *Apriori*, para que, a cada ciclo, as regras candidatas sejam geradas. Esses ciclos são aproveitados como pontos de referência na visualização das regras que estão sendo mineradas e no ajuste das medidas de interesse e dos parâmetros de entrada. Dessa forma, o resultado intermediário da mineração pode ser analisado e, se for o caso, as medidas e os parâmetros são redefinidos para realimentar o algoritmo. O serviço foi utilizado no exemplo de aplicação descrito no Capítulo 7.

### 5.1.3 MEDIDAS DE INTERESSE E PARÂMETROS DE ENTRADA

As medidas de interesse e os parâmetros de entrada são aplicados para filtrar o que continuará no processo de mineração, dessa forma, nem todas as regras geradas em um ciclo são passadas para o seguinte. A principal medida de interesse utilizada para definir se uma regra será descartada é o suporte mínimo. As regras encontradas em um ciclo permanecem na mineração, apenas, se atenderem o critério de suporte mínimo especificado pelo usuário. Outras medidas de interesse também podem ser aplicadas. O Serviço de Geração de Regras de Associação (seção 5.1.2), utilizado no exemplo de aplicação do Capítulo 7, é alimentado com as seguintes medidas de interesse para minerar regras:

- `minSuporte`: suporte mínimo que uma regra deve possuir;
- `minConfiança`: confiança mínima que uma regra deve possuir;
- `minLift`: *lift* mínimo que uma regra deve possuir; e
- `maxConvicção`: convicção máxima que uma regra deve possuir.

Outro mecanismo utilizado por Lavôr (2003) para reduzir o número de regras a cada ciclo é a definição de regras que são triviais. As regras triviais são descartadas no próximo ciclo. Do mesmo modo, é possível definir regras que são importantes para que elas permaneçam no processo, independente das medidas de interesse. Os seguintes parâmetros são utilizados para definir as regras que devem ser incluídas ou excluídas da mineração:

- `ArquivoDefCombTrivial`: caminho e nome completo do arquivo texto com a especificação das combinações e regras triviais a serem eliminadas pelo serviço; e
- `ArquivoDefCombImprescindívies`: caminho e nome completo do arquivo texto com a especificação das combinações que são imprescindíveis e devem ser mantidas pelo serviço.

Além dos parâmetros citados, o Serviço de Geração de Regras de Associação (LAVÔR, 2003) utiliza os seguintes parâmetros durante a mineração:

- `FatorAmostra`: percentual de registros do `ArquivoDados` que deve ser extraído para a composição da amostra;
- `numMaxItens`: número máximo de itens em um conjunto de itens freqüentes;
- `numMaxItensAntec`: número máximo de itens permitido no antecedente de uma regra;
- `numMaxItensConseq`: número máximo de itens permitido no conseqüente de uma regra; e
- `numMaxRegras`: número máximo de regras a serem geradas pelo serviço.

### 5.1.4 FERRAMENTA OLAP

As regras de associação resultantes do processo de mineração são armazenadas no ambiente analítico proposto, que tem como característica a não-volatilidade, ou seja, é um ambiente essencialmente histórico, onde o resultado de todo o processo de exploração é guardado ao longo do tempo. Uma ferramenta OLAP permite que toda a memória da mineração possa ser analisada. No exemplo de aplicação apresentado no Capítulo 7, o *Microsoft SQL Server for Analisys Services* foi utilizado para a navegação pelas regras mineradas.

### 5.1.5 REGRAS DE ASSOCIAÇÃO E INFORMAÇÕES DE CONTROLE

As regras de associação geradas em cada um dos ciclos do algoritmo *Apriori*, que vai de 1 até N, são armazenadas em um cubo no ambiente analítico proposto, conforme o modelo de dados multidimensional definido no item 6.2. Para cada regra, são armazenadas as informações de controle utilizadas durante o processo de sua mineração, tais como: o ciclo, o valor de suas medidas de interesse, se foi descartada ou aproveitada, e a que carga do repositório de dados pertence.

O resultado de todo o processo analítico de exploração é gerado após o último ciclo da mineração. Nessa etapa, todas as regras de associação que não foram descartadas durante o processo são armazenadas no ambiente analítico para serem avaliadas pelo usuário.

## 5.2 INTEGRAÇÃO DAS OPERAÇÕES ANALÍTICAS E DE MINERAÇÃO

De acordo com os conceitos definidos em (HAN, 1997), a integração das funcionalidades analíticas com a mineração de regras de associação aqui definida é do tipo *cubing while mining with backtracking*. Esse processo é constituído pelos seguintes passos, ilustrados na Figura 8 e descritos a seguir: (i) *cubing*: montar o cubo a ser minerado, a partir de uma porção do repositório de dados e de um nível de granularidade definidos pelo usuário; (ii) *mining*: alimentar o Serviço de Mineração de Regras de Associação com o cubo montado; (iii) *cubing while mining*: carregar o cubo das regras de associação com o resultado de cada ciclo do algoritmo para análise do usuário; (iv) *backtracking*: calibrar as medidas de interesse e parâmetros de entrada, depois da análise do resultado do passo anterior, para eventuais redirecionamentos no caminho da mineração; e (v) *cubing*: armazenar as regras de associação mineradas em um cubo para análise, manipulação e validação.

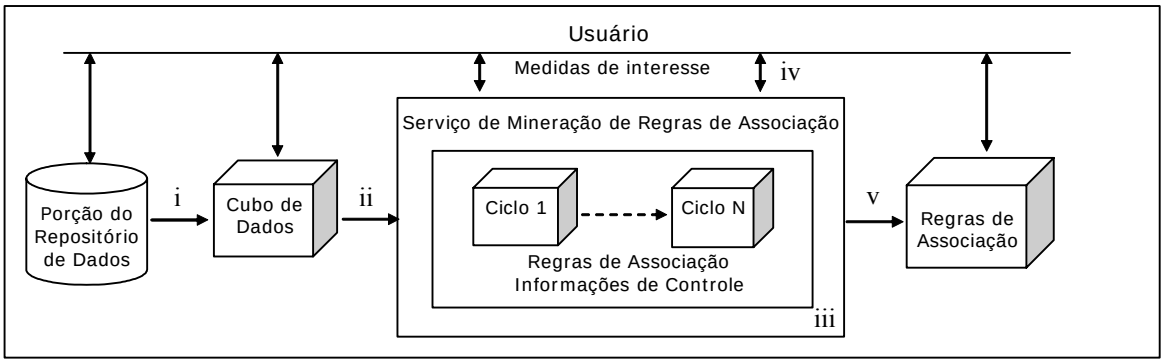



**Figura 8: Integração das operações analíticas com a mineração de regras de associação**

## 5.3 PROCESSO ANALÍTICO DE EXPLORAÇÃO DE REGRAS DE ASSOCIAÇÃO

Todos os elementos da arquitetura proposta foram empregados, sistematicamente, em um processo analítico de exploração de regras de associação, o qual visa a descoberta do conhecimento. A interferência do usuário a cada ciclo da mineração de regras é inserida no processo de descoberta do conhecimento (Capítulo 2). As funcionalidades analíticas de uma ferramenta OLAP se integram ao processo ao apresentar os resultados de cada ciclo da mineração. As outras etapas que compõem esse processo são (Figura 9): seleção da porção do repositório de dados que será minerada, transformação dessa porção para o formato de dados utilizado pelo serviço de mineração, apresentação dos padrões de dados encontrados, e avaliação e análise das regras de associação pelo usuário final.

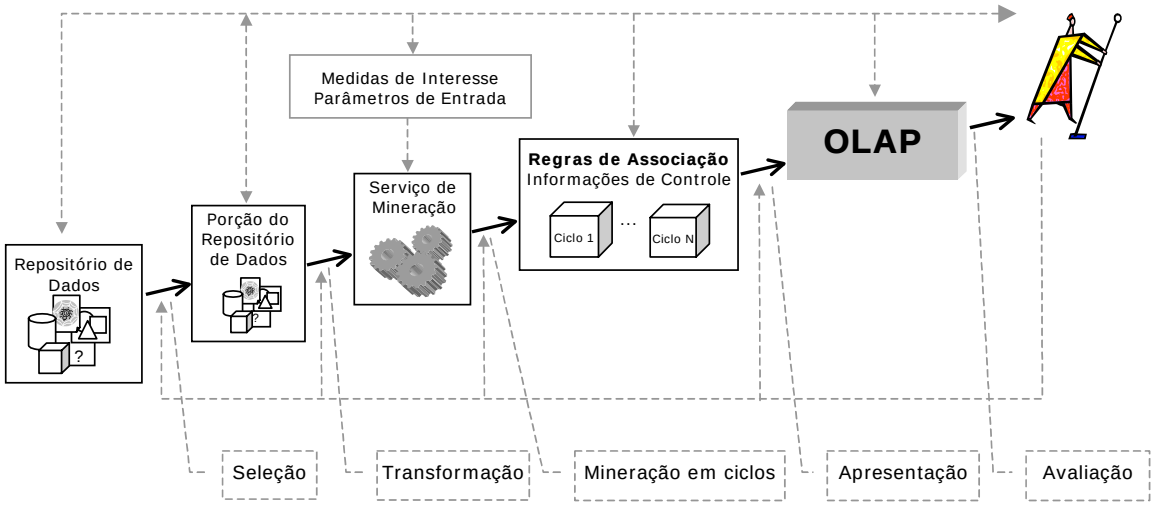



**Figura 9: Análise das Regras de Associação como mais um passo no Processo de Descoberta do Conhecimento**

A mineração de regras de associação é realizada por meio de ciclos. Em cada ciclo as regras candidatas são geradas e armazenadas em um cubo para serem analisadas pelo usuário, através de uma ferramenta OLAP. Assim, como o processo de descoberta do conhecimento é uma seqüência iterativa de etapas, a mineração de regras de associação proposta também é

realizada por meio de passos repetitivos, onde cada passo é um ciclo do algoritmo. As regras geradas ao final de cada ciclo são disponibilizadas para serem analisadas pelo usuário. A partir dessa análise, o usuário percebe que algumas regras interessantes foram descartadas e que regras não interessantes constam no resultado da mineração. Com isso, o usuário pode redefinir as medidas de interesse e os parâmetros de entrada. Caso haja uma nova definição, o mesmo ciclo é executado, gerando novos resultados.

A interação do usuário com o serviço de mineração é feita por meio da definição de medidas de interesse e os parâmetros de entrada que podem ser redefinidos a cada iteração, permitindo a interferência na exploração das regras de associação. No processo apresentado, o usuário pode consultar o repositório de dados original ou a porção dele para, junto com o conhecimento adquirido, redefinir as medidas de interesse e os parâmetros de entrada.

Na Figura 9, as setas pontilhadas representam o fluxo de interação entre os elementos do processo. Observa-se que o usuário interage com todos os elementos do processo de descoberta do conhecimento, a qualquer momento e em todas as etapas, destacando-se a interferência no processo de geração de regras de associação por meio da calibragem das medidas de interesse e dos parâmetros de entrada que alimentam o serviço de mineração.

Observa-se que as funcionalidades analíticas apresentadas visam contribuir para a redução do número de elementos do conjunto de regras no final do processo de exploração. Espera-se que, nesse conjunto reduzido, estejam contidas as regras mais interessantes e, assim, conhecimento novo possa ser descoberto.

# 6 UTILIZAÇÃO DA ARQUITETURA PROPOSTA

Este capítulo apresenta como os elementos da arquitetura de exploração de regras de associação em um ambiente analítico podem ser utilizados, de forma integrada, para interferir no processo de mineração, e como a memória de todo o processo pode ser armazenada em um cubo.

A interferência no processo de mineração aqui proposta é realizada por meio da redefinição das medidas de interesse e dos parâmetros de entrada pelo usuário após analisar a memória da mineração, que contém os resultados intermediários. Esses resultados são armazenados no formato de um cubo em um banco de dados multidimensional para serem analisados através de uma ferramenta OLAP. Os passos da interferência estão descritos na próxima seção.

O cubo da memória do processo de mineração contém, além das informações de controle do processo, as regras geradas a cada ciclo do algoritmo *Apriori*. Esse cubo contém os itens freqüentes e uma marcação para identificar as regras que passarão para o próximo ciclo e quais serão descartadas. Dentre as regras excluídas, o usuário pode selecionar aquelas que devem permanecer no processo. Pode, também, alterar o suporte mínimo de acordo com a quantidade de regras geradas, além de outros ajustes nos parâmetros de entrada. O cubo fornece apoio ao processo de exploração e depois à escolha das regras de associação mais interessantes. O modelo de dados multidimensional da memória do processo de exploração das regras de associação é apresentado na seção 6.2.

## 6.1 INTERFERÊNCIA NO PROCESSO DE EXPLORAÇÃO

Soluções anteriores, como visto na seção 3.4, reduzem a quantidade de regras de associação geradas, porém de forma artificial. Esse procedimento pode excluir regras relevantes do resultado final. Para evitar isso, este trabalho propõe que o algoritmo mostre o que está sendo descoberto e, a partir daí, o usuário pode interferir no processo indicando se interessa ou não. Sistemas de mineração de dados devem permitir que o usuário especifique dicas para guiar ou focar a busca por padrões interessantes (HAN e KAMBER, 2001).

A influência nas decisões tomadas sobre o caminho que o algoritmo está percorrendo para a mineração é realizada por meio da alimentação de novos parâmetros no momento em que o algoritmo toma decisões. Tais decisões são, normalmente, tomadas a partir de dados previamente definidos no início da mineração. No ambiente analítico proposto, as decisões são efetuadas a partir das informações introduzidas pelos usuários durante todo o processo. Quanto maior o conhecimento do usuário sobre o negócio, melhor será o direcionamento da mineração e regras mais pertinentes serão geradas.

A interferência no processo analítico de exploração de regras de associação está ilustrada no Diagrama de Atividades da Figura 10 e ocorre de acordo com os seguintes passos:

1. O usuário inicia o processo definindo a porção do repositório de dados na qual será realizada a mineração. Essa porção é extraída do repositório de dados do ambiente transacional ou do *data warehouse* do ambiente analítico. Durante o processamento, o usuário analisa o caminho que a exploração está percorrendo e, a partir daí, a porção do repositório pode ser redefinida. Essa redefinição é feita para incorporar dados que foram deixados de fora, e que podem ajudar na busca por informações úteis, ou para retirar do processo os dados que não trazem conhecimentos novos. Na mineração dos dados do vestibular de 2004 da

UFRJ, descrita no Capítulo 7, percebeu-se que a Unidade da Federação (UF) do candidato não contribuía na busca de conhecimentos novos sobre o perfil da família[5] do vestibulando, e que se os dados dos vestibulandos de 2003 fossem acrescentados ao processo de mineração, as regras de associação geradas seriam mais precisas. A partir dessa percepção, a UF foi retirada do processamento e os dados do ano de 2003 foram incorporados ao processo.

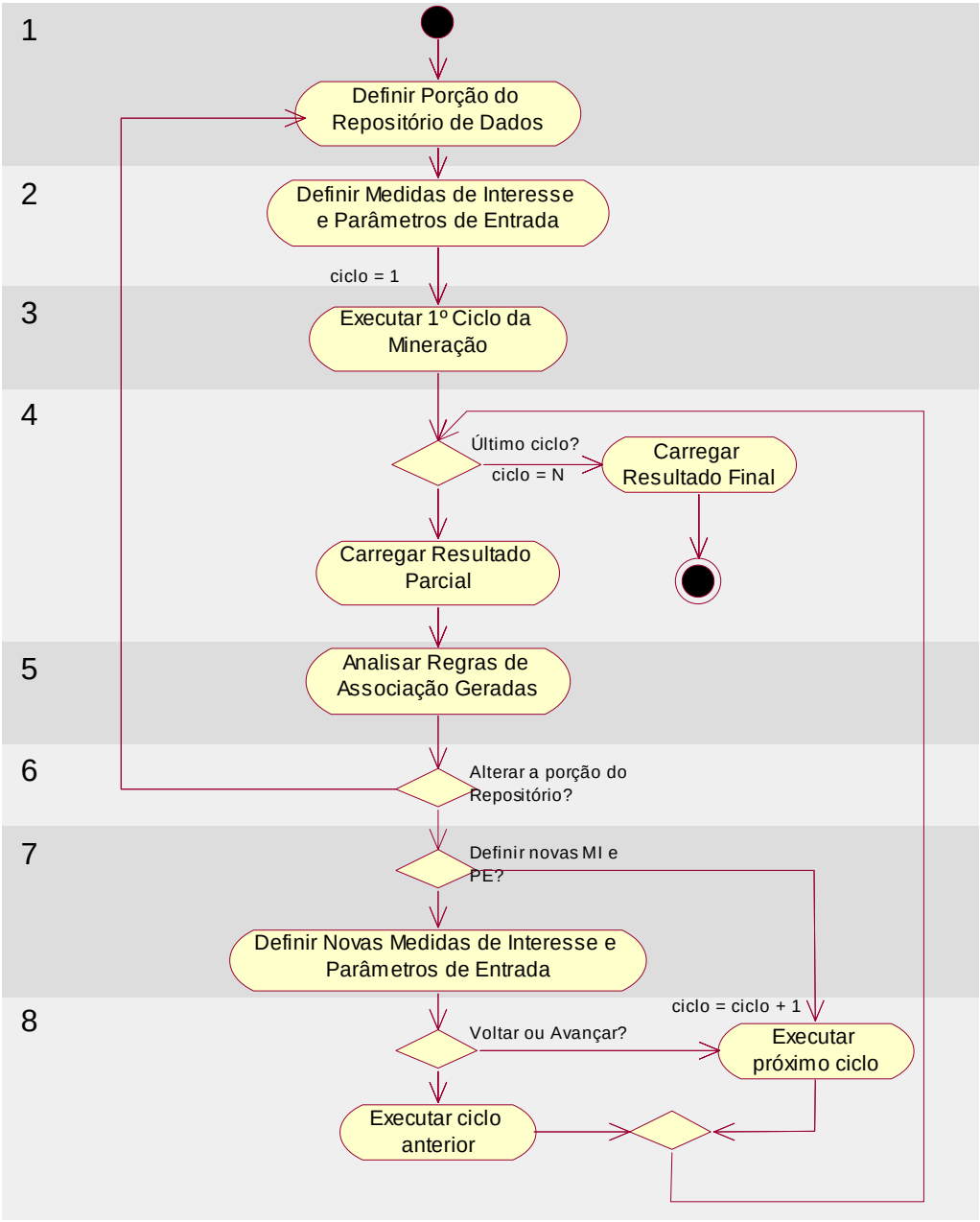



**Figura 10: Interferindo no Processo Analítico de Exploração de Regras de Associação**

[5] As informações sobre a família do candidato constam no questionário sócio-cultural preenchido no ato da inscrição, que possui informações sobre a composição da família, renda, quantidade de cômodos da casa, automóveis e livros, nível de instrução dos pais, entre outros.

2. No segundo passo do processo, o usuário define as medidas de interesse e os parâmetros de entrada que devem constar ou não no resultado da execução do primeiro ciclo da mineração. Para maiores informações sobre a composição desses parâmetros, ver seção 5.1.3. Essas informações são baseadas no conhecimento que o usuário já possui sobre o negócio, sobre o assunto que ele está querendo obter novos conhecimentos e sobre a porção do repositório de informações que ele definiu. No exemplo de aplicação do Capítulo 7, foram feitas as seguintes definições: o número máximo de itens no antecedente igual a dois e no conseqüente igual três; a regra `situaçãoFinal|reprovado>>matriculadoUFRJ|não` é trivial e não deve constar no resultado da mineração, devendo ser excluída do processamento; e as regras que ocorrem em mais de 70% das transações devem ser mineradas, para isso o suporte mínimo foi definido em 70%.

3. Com a porção do repositório de dados e as informações de entrada definidas, o Serviço de Mineração de Regras de Associação está pronto para iniciar o processamento. O primeiro passo do algoritmo *Apriori* (primeira iteração) gera os itens freqüentes de acordo com o suporte mínimo especificado. Depois disso, o primeiro ciclo da mineração é executado. No exemplo de aplicação do Capítulo 7, uma janela com uma barra de progresso é exibida, mostrando o andamento da mineração. O serviço, inicialmente, gera a amostra dos dados que serão minerados de acordo com o parâmetro de entrada `fatorAmostra`. A seguir, as regras de associação candidatas com apenas dois itens freqüentes (atendem ao suporte mínimo, mas não aos outros parâmetros) são geradas na tabela candidata e armazenadas em um arquivo do tipo texto. Essas regras são o resultado do primeiro ciclo da mineração.

4. No quarto passo do processo analítico de exploração, a tabela candidata (seção 3.5.3) é verificada para determinar se haverá um próximo ciclo ou não. O último ciclo do algoritmo *Apriori* ocorre quando a tabela candidata está vazia. Nesse momento, as regras de associação finais são geradas e o resultado final da mineração é carregado no ambiente

analítico. Caso a tabela candidata não esteja vazia, um próximo ciclo é executado. Nesse momento, as regras são extraídas da tabela candidata, carregadas como resultado parcial da mineração e disponibilizadas para avaliação e análise do usuário. No exemplo da aplicação, por limitações de tempo e escopo, o Serviço de Mineração de Regras de Associação gera, de uma vez só, todos os arquivos com as regras candidatas de cada ciclo e os arquivos com as regras que foram aproveitadas para o ciclo seguinte. De acordo com a arquitetura proposta, os arquivos gerados são transformados e carregados em um banco de dados. No exemplo de aplicação, as tabelas do modelo de dados multidimensional, descritas na seção 6.2, são criadas em um banco de dados relacional e alimentadas com a memória da mineração contida nos arquivos gerados pelo Serviço de Mineração de Regras de Associação.

5. Com o ambiente analítico carregado, o usuário pode analisar o resultado daquele ciclo da mineração, confrontar os dados com o repositório de informações original e com o conhecimento previamente adquirido através da sua experiência do negócio e decidir se deseja interferir no processo da mineração. Como pode ser observado no exemplo de aplicação, os arquivos tipo texto gerados não são legíveis e são difíceis de serem manipulados, prejudicando a análise. A solução desse problema, proposta na arquitetura, é montar um cubo a partir do banco de dados carregado. A ferramenta OLAP é utilizada para possibilitar a navegação pela memória da mineração e a sua análise sob diversas perspectivas.

6. Durante as análises do passo anterior, o usuário pode perceber que algumas informações de seu interesse não constam no resultado do ciclo e nem na porção do repositório de dados que está sendo minerada. Nesse momento, o usuário pode interferir no processo, voltando ao primeiro passo para redefinir a porção do repositório de informações original a ser minerada, reiniciando todo o processo. No exemplo de aplicação, inicialmente, apenas os dados de 2004 estavam sendo minerados. Percebeu-se que os dados de 2003 eram importantes para o processo de descoberta do conhecimento sobre o perfil dos vestibulandos.

Dessa forma, o usuário interferiu no processo, redefinindo a porção do repositório incluindo os dados de 2003.

7. Uma vez que todas as informações relevantes tenham sido selecionadas, o usuário pode passar para o próximo ciclo da mineração ou pode, novamente, interferir no processo por meio da definição de novas medidas de interesse e parâmetros de entrada. É importante ressaltar que o suporte mínimo só pode ser alterado para um valor superior ao inicialmente especificado. Como foi visto no item 3.5, o primeiro passo do algoritmo *Apriori* gera a tabela de itens freqüentes utilizando como parâmetro o suporte mínimo, e todas as iterações seguintes são realizadas utilizando esta tabela. Dessa forma, para alterar o suporte mínimo da mineração é necessário que o processo seja reiniciado.

8. Nesse passo, mais um ponto de interferência no processo analítico de exploração de regras de associação é oferecido. O usuário decide se as medidas redefinidas no passo anterior serão utilizadas para executar, novamente, o mesmo ciclo ou o próximo. Essa etapa é executada incrementando ou decrementando a variável *k* do algoritmo *Apriori* (seção 3.5), isto é, adicionando um item ao conjunto de itens freqüentes.

Após a realização de todas as etapas, o processo retorna ao quarto passo até que o último ciclo do algoritmo seja executado. Ao término do processo, o resultado final é carregado no *data mart*, onde as regras de associação geradas poderão ser analisadas pelo usuário, e se for o caso, o processo poderá ser reiniciado com novas informações de entrada.

Uma outra forma de apresentação do processo analítico de exploração está ilustrada na Figura 11, onde as atividades dos diferentes agentes que atuam no processo são posicionadas em raias (*swinlines*). Os agentes que atuam no processo são: Usuário, Serviço de Mineração de Regras de Associação e Banco de Dados.

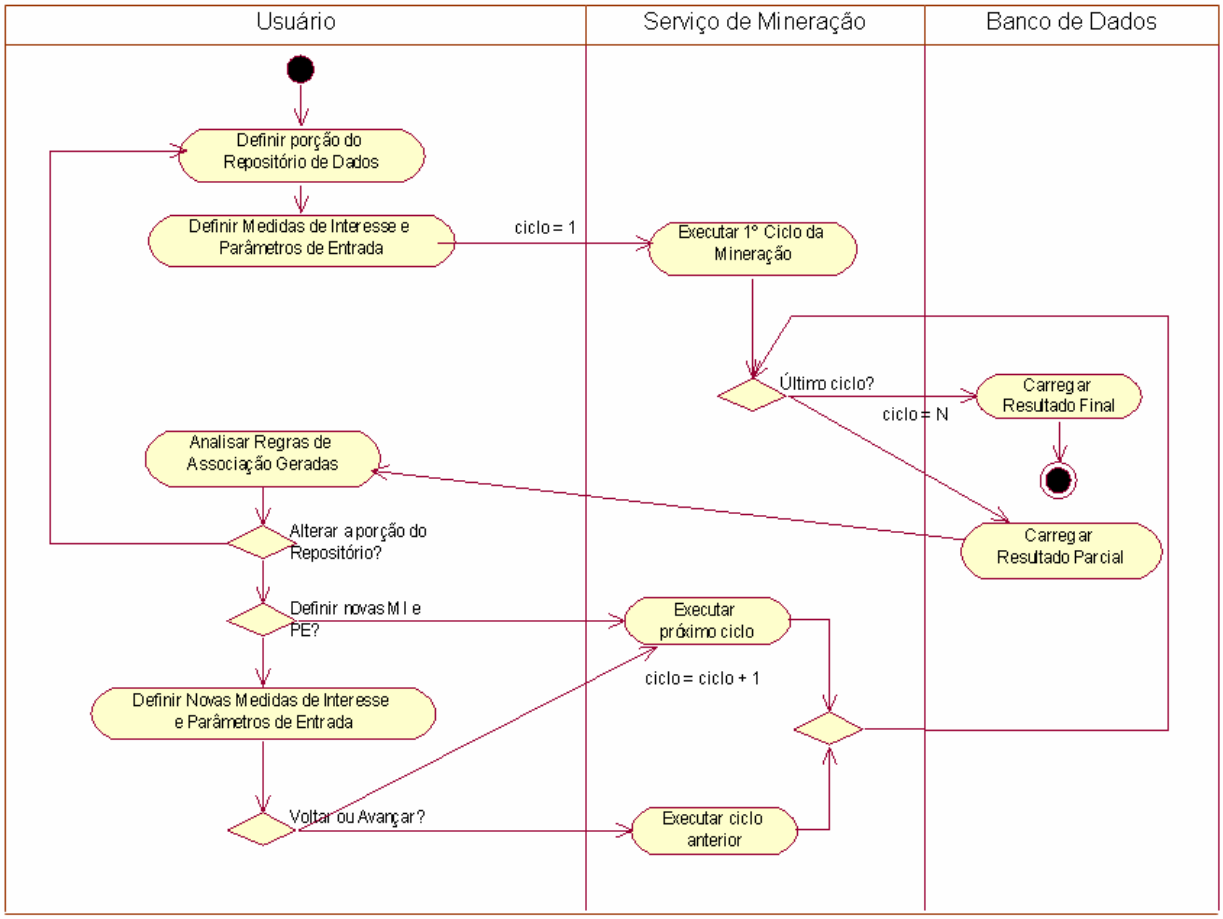



**Figura 11: Papéis do Processo Analítico de Exploração de Regras de Associação**

## 6.2 MODELO DE DADOS MULTIDIMENSIONAL DE REGRAS DE ASSOCIAÇÃO

Nesta seção é apresentado o modelo de dados multidimensional utilizado na montagem do cubo do processo de exploração de regras de associação, que será carregado na ferramenta OLAP. Esse cubo é constituído pelas regras de associação e suas medidas de interesse, o ciclo a que cada regra pertence, uma marcação dizendo se a regra é aproveitada ou descartada para o ciclo seguinte, uma referência ao estado do banco de dados original no momento da mineração, e os parâmetros de entrada utilizados para aquele processamento. A Figura 12 ilustra o esquema floco de neve (seção 2.1.3) do modelo de dados multidimensional do processo analítico de exploração de regras de associação.

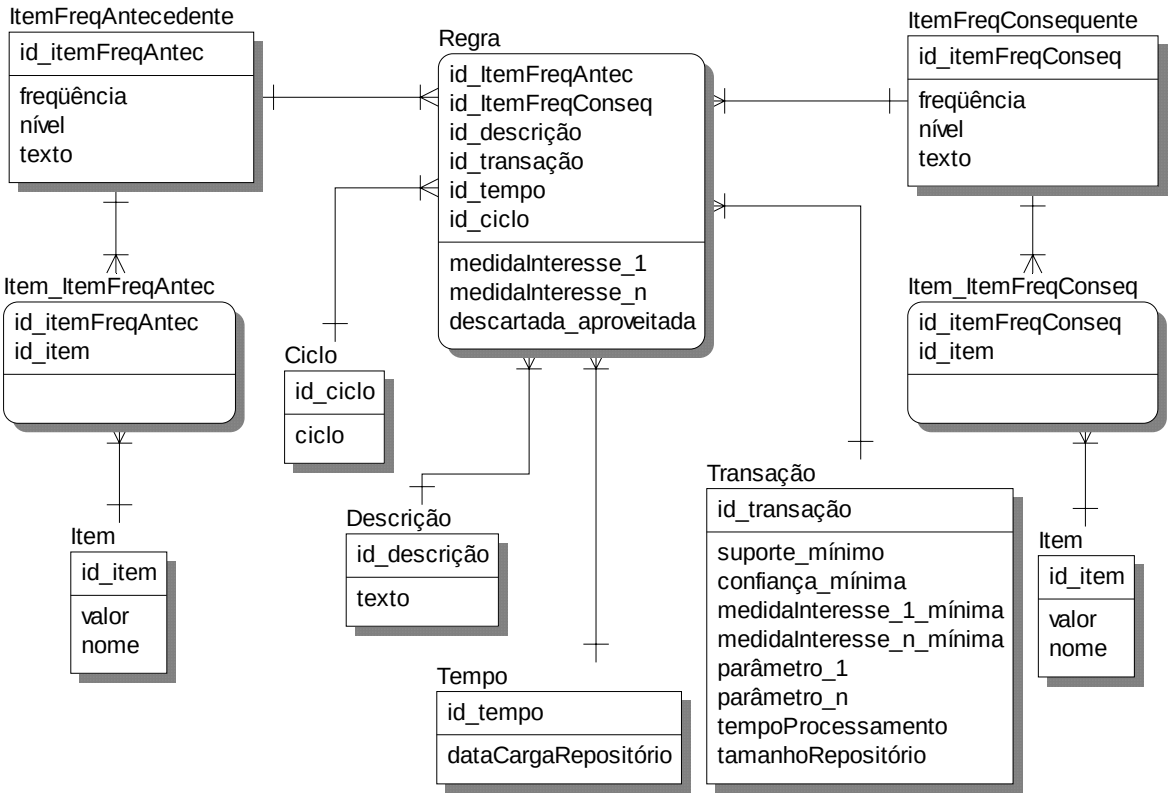



**Figura 12: Modelo de Dados Multidimensional de Regras de Associação**

O fato e as dimensões que compõem o modelo multidimensional do processo de exploração de regras de associação são apresentados nas próximas subseções.

### 6.2.1 FATO

Em geral, as consultas analíticas são feitas para extrair dados agregados que irão compor um relatório estruturado a ser analisado. O modelo de dados multidimensional do processo de mineração de regras de associação contém fatos não-aditivos, ou seja, operações aritméticas não podem ser realizadas nas suas métricas, pois resultariam em dados sem significado. A regra de associação é considerada como o fato desse modelo, e as métricas são as medidas de interesse e um atributo do tipo sim/não (*flag*) utilizado para representar se a regra foi descartada ou aproveitada naquele ciclo.

Uma outra abordagem para a modelagem da tabela fato de regras de associação é a utilização de uma tabela de cobertura, ou tabela fato sem fatos (*factless fact table*) (KIMBALL et al., 1998), com todas as combinações possíveis entre todos os possíveis itens freqüentes, onde as medidas de interesse calculadas indicam a significância da regra. Essa abordagem pode ser considerada para porções de dados a serem minerados que possuam uma pequena quantidade de itens em seu repositório.

### 6.2.2 DIMENSÕES

As dimensões definidas para o modelo de dados multidimensional do processo de exploração são: Tempo, Ciclo, Item Freqüente, Item, Item_ItemFreqüente, Transação e Descrição.

#### 6.2.2.1 DIMENSÃO TEMPO

A dimensão tempo desempenha um papel fundamental no ambiente analítico proposto, pois o resultado do processo de mineração para cada estado do banco de dados será armazenado, o que permitirá uma análise da mudança das regras de associação ao longo do tempo. A dimensão tempo é formada por um atributo com a data da carga do repositório de dados, ou seja, a data em que os dados foram extraídos de sua origem para serem minerados. Esse atributo funciona como um identificador da extração do repositório, identifica o momento ao qual aquele estado do banco pertence.

No *data warehouse* tradicional, a dimensão tempo permite a análise do comportamento dos fatos ao longo do tempo em que ocorreram. No *data mart* de regras de associação, a dimensão tempo permite a análise das regras ao longo das cargas do repositório de dados. A partir de uma carga executada em um determinando momento no tempo, é possível a realização de várias transações de mineração. Cada transação gera várias regras que alimentam a tabela de fatos. Por exemplo, os dados extraídos do ambiente transacional no início do ano podem ser minerados durante todo o primeiro semestre, e muitas regras podem ser geradas. No segundo semestre, uma nova extração é realizada e novas minerações podem ser executadas. O resultado da mineração do primeiro semestre pode ser comparado com o resultado da mineração do segundo semestre, revelando a mudança das regras ao longo do ano.

#### 6.2.2.2 DIMENSÃO CICLO

A dimensão ciclo contém um atributo que armazena o ciclo ao qual a regra pertence. Esse atributo permite a análise do caminho percorrido pela regra indicando em que ciclo foi gerada e em que ciclo foi descartada do processamento. Com isso, é possível a determinação de novas medidas de interesse e parâmetros de entrada para que determinadas regras sejam

excluídas logo no início do processamento, ou para que permaneçam até o final. Permite, também, a análise da quantidade de regras geradas e descartadas a cada ciclo.

#### 6.2.2.3 DIMENSÃO ITEM FREQÜENTE

O item freqüente aparece no modelo como uma dimensão desempenhando dois papéis, antecedente e conseqüente. Essa dimensão é denominada *Role Playing Dimension* por Kimball (1998).

A dimensão item freqüente contém: o nível ao qual o item freqüente pertence, ou seja, o número de itens que o compõe; a freqüência, isto é, o suporte do item; e um texto que descreve o item para fornecer semântica às consultas analíticas.

#### 6.2.2.4 DIMENSÃO ITEM

Cada item de um conjunto de itens freqüentes é referenciado como uma dimensão por Han e Kamber (2001). Uma regra é composta por vários itens, e com isso recebe a denominação de regra de associação multidimensional.

Como um item freqüente é composto por um ou mais itens, e cada item pode pertencer a um ou mais itens freqüentes, a dimensão Item foi modelada contendo um relacionamento de muitos-para-muitos com a dimensão Item Freqüente. Essa dimensão contém como atributos o nome e o valor do(s) item(ns) que compõe(m) o item freqüente antecedente ou conseqüente de uma regra.

#### 6.2.2.5 DIMENSÃO ITEM_ITEMFREQÜENTE

A dimensão Item_ItemFreqüente é do tipo dimensão ponte (*Bridge Table*) ou dimensão de navegação (*Navigation Table*) (KIMBALL et al., 1998) e é utilizada para resolver o problema das dimensões muitos-para-muitos, como é o caso das dimensões Item e Item Freqüente. Esse tipo de dimensão também é chamado tabela fato sem fatos (*Factless*

*Fact Table*) (KIMBALL et al., 1998), pois representa um fato entre as dimensões que ela interliga, sem possuir medidas.

#### 6.2.2.6 DIMENSÃO TRANSAÇÃO

A dimensão transação contém as informações sobre o processamento da regra: os parâmetros de entrada e as medidas de interesse utilizados pelo serviço, o tempo gasto para a mineração e o tamanho do banco de dados. As características da transação que gerou a regra podem ajudar o usuário, no momento das análises das regras, na calibragem das medidas de interesse e dos parâmetros de entrada. Por exemplo, suponha que o usuário observou que o serviço de mineração gastou muito tempo para gerar as regras, mas deseja aumentar o número máximo de itens freqüentes da regra e diminuir o tempo de mineração. Para isso, analisa a quantidade de regras e as características da transação que as gerou e decide, então, diminuir o fator de amostra, o que irá acarretar na redução do tempo gasto pelo serviço.

#### 6.2.2.7 DIMENSÃO DESCRIÇÃO

A dimensão descrição não é considerada como mais uma perspectiva na análise das regras de associação. Entretanto, é utilizada para facilitar a leitura das regras pelo usuário. O atributo texto armazena a descrição textual da regra. Esse atributo poderia estar modelado na tabela fato, mas por questões de projeto foi modelada em uma dimensão separada.

### 6.2.3 PERGUNTAS ANALÍTICAS DE APOIO À EXPLORAÇÃO DE REGRAS

O modelo de dados proposto visa responder perguntas analíticas por meio dos relatórios oferecidos pelas ferramentas OLAP. Alguns desses relatórios estão ilustrados nos exemplos da última etapa do processo analítico de exploração, apresentados no Capítulo 7. A seguir estão exemplos de perguntas analíticas:

- Quais regras desaparecem ao longo do processamento? (Figura 27)
- Quais regras têm um determinado item freqüente no seu conseqüente? (Figura 29)
- Qual o impacto de uma determinada medida de interesse?
- Quais itens freqüentes dependem de um determinado parâmetro?
- Quais regras têm o item freqüente x como antecedente e o item freqüente y como conseqüente, com todos os valores de x e y? (Figura 31)

Além de responder perguntas, o modelo pode mostrar informações para serem analisadas e permitir navegações, como descrito a seguir:

- Quantidade de regras aproveitadas e descartadas por ciclo e por transação;
- A partir de uma carga do repositório navegar para todas as transações executadas naquela carga;
- Navegação de cada ciclo para a situação das regras, aproveitadas ou descartadas.
- Quantidade de regras geradas por transação dentro de cada carga e por ciclo;
- Quantidade de regras em uma determinada faixa de valores de uma medida de interesse;
- Navegação por todos os níveis dos itens freqüentes e das regras;
- Quantidade de regras por nível do item freqüente (Figura 32);
- Quantidade de regras que têm no seu antecedente o item freqüente x com todos os seus valores; e
- Quantidade de regras geradas a cada ciclo de acordo com as características da transação (Figura 33).

# 7 EXEMPLO DE APLICAÇÃO

Algumas etapas do processo analítico de exploração de regras de associação foram implementadas utilizando os componentes da arquitetura proposta. O objetivo do exemplo de aplicação é demonstrar a utilização da arquitetura e do processo analítico de exploração como uma estratégia viável no processo de descoberta de conhecimento.

Na aplicação, foi utilizado como repositório de dados o banco de dados dos candidatos ao vestibular da UFRJ dos anos 2003 e 2004, juntamente com os dados de um questionário sócio-cultural respondido pelos candidatos no ato da inscrição. Esse banco de dados contém informações pessoais (ex.: sexo, endereço, estado civil e naturalidade); informações sócio-culturais (ex.: nível de instrução do pai e da mãe, quantidade de cômodos da casa e situação financeira da família); informações do vestibular (ex.: aprovado ou reprovado, se foi matriculado ou não na UFRJ e as notas de cada disciplina no vestibular). Os campos desse banco de dados respondem perguntas independentes, o que contribui para a busca de itens freqüentes.

No processo analítico de exploração de regras de associação, alguns atributos do repositório de dados são selecionados pelo usuário, extraídos e transformados para alimentar o Serviço de Mineração de Regras de Associação. O serviço gera as regras utilizando as restrições definidas pelo usuário por meio das medidas de interesse e dos parâmetros de entrada. As regras geradas e as informações de controle são armazenadas no formato de um cubo, para serem apresentadas ao usuário através de uma ferramenta OLAP. Em qualquer uma das etapas, o usuário pode interferir no processo e direcionar a exploração para as regras mais interessantes. Todas essas etapas fazem parte do processo analítico de exploração de regras de associação descrito na seção 5.3.

A infra-estrutura utilizada no exemplo de aplicação é composta dos seguintes recursos (Tabela 2): um computador, um ambiente de desenvolvimento, um banco de dados, uma ferramenta de integração de dados, um serviço de mineração de regras e uma ferramenta OLAP. A Tabela 2 apresenta, também, as especificações utilizadas na implementação do exemplo para cada um dos recursos. Entretanto, os recursos não precisam, necessariamente, obedecer a essas especificações, podendo ser utilizados recursos de diferentes fabricantes.

**Tabela 2: Infra-estrutura do exemplo de aplicação**

| Recurso | Especificação |
|---|---|
| Computador | *Pentium* IV 3.2 GHz *Hyper Thread*, 1GB de Memória RAM *Dual Chanel* e HD 40GB de 7200 RPM. |
| Ambiente de desenvolvimento | *Microsoft Development Environment 2003 - Visual Basic .Net Project* |
| Banco de dados | *Microsoft SQL Server 2005* |
| Integração de dados (leitura, transformação e carga) | *Microsoft Visual Studio 2005 - Business Intelligence Projects – Integration Services Project* |
| Serviço de Mineração de Regras de Associação | Serviço de Geração de Regras de Associação[6] (LAVÔR, 2003) |
| Ferramenta OLAP | *Microsoft Visual Studio 2005 - Business Intelligence Projects – Analysis Services Project* |

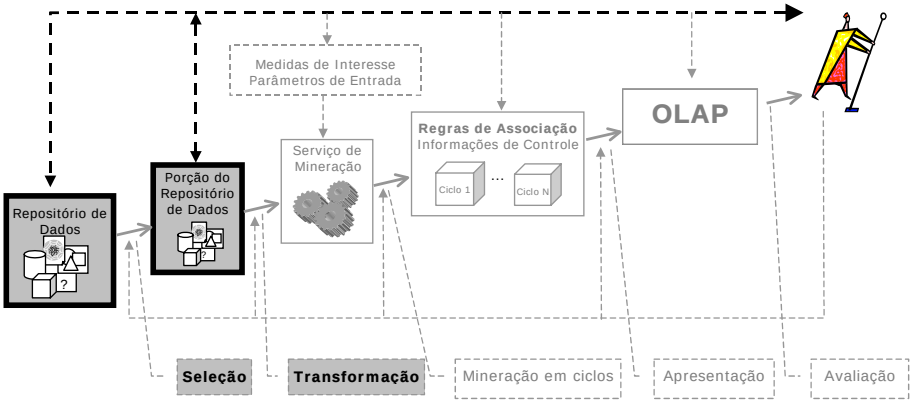
Durante a execução do exemplo de aplicação, observou-se que o ambiente analítico facilitou a navegação pelas regras descobertas a cada ciclo da mineração. A partir das análises realizadas durante a navegação, novas medidas de interesse e novos parâmetros de entrada foram definidos e realimentaram o Serviço de Mineração de Regras de Associação, direcionando a exploração para dados mais relevantes.

[6] Conforme dito no item 5.1.2, o Serviço de Geração de Regras de Associação (LAVÔR, 2003) foi adaptado às necessidades do ambiente analítico proposto.

Durante a aplicação do processo analítico de exploração de regras de associação, observou-se que a arquitetura proposta é viável, todas as etapas foram executadas, os dados foram transformados e minerados, as regras e as informações sobre o processo foram apresentados ao usuário, permitindo a análise de todo o caminho da mineração. Com isso, essa arquitetura permite uma exploração flexível e centrada no humano, o que contribui para a descoberta de regras de associação mais interessantes.

Será apresentada, a seguir, a descrição da implementação das etapas do processo analítico de descoberta do conhecimento utilizadas no exemplo de aplicação. No início da descrição de cada uma das etapas, está ilustrado o processo analítico de descoberta do conhecimento, com um destaque em negrito, da etapa que está sendo descrita.

## 7.1 ETAPA 1 – DEFINIÇÃO DA PORÇÃO DO BANCO DE DADOS




O processo analítico de descoberta do conhecimento é iniciado com a Seleção, pelo usuário, da porção do repositório de dados original na qual será realizada a mineração. O repositório de dados, usado no exemplo, está armazenado em um banco de dados do *Microsoft Access* e contém as seguintes tabelas: `VEST_Candidato` com 100.528 registros e 30 atributos, `VEST_SocioCultural` com 100.403 registros e 42 atributos, e `VEST_Nota`

com 569.000 registros e 6 atributos. O modelo de dados do repositório é apresentado na Figura 13.

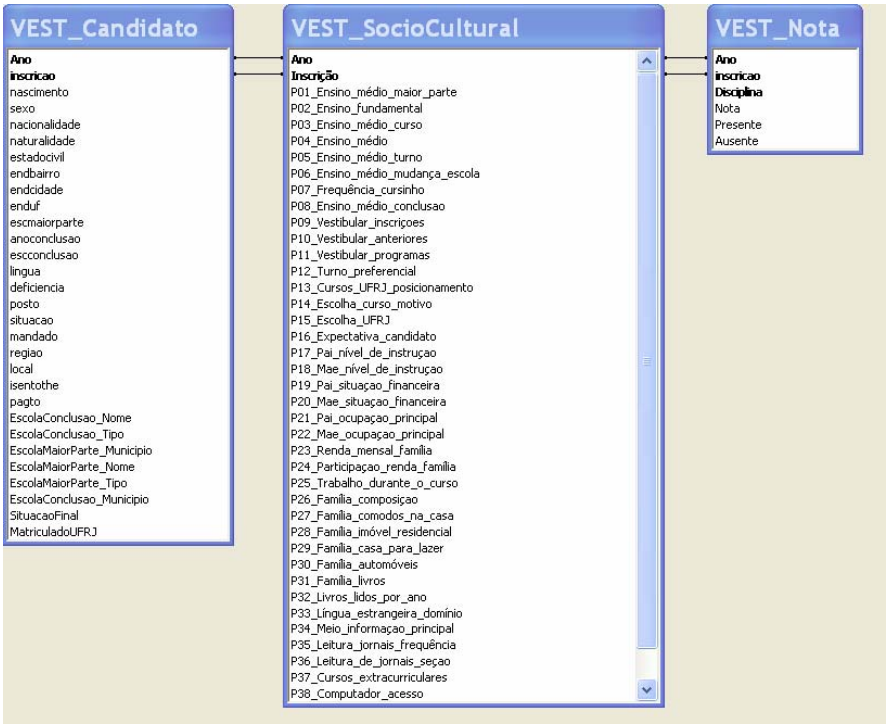



**Figura 13: Modelo de Dados do Repositório**

Após a seleção da porção do repositório de dados que será minerada, esses dados são extraídos e **Transformados** para um formato de um arquivo texto compreendido pelo Serviço de Mineração de Regras de Associação. Nesse formato, não há separação entre os campos e cada linha do arquivo representa um registro, conforme ilustrado na Figura 14. O arquivo extraído contém 53.213 registros, cada um com 35 atributos, e ocupa 72,8 MB de espaço em disco.

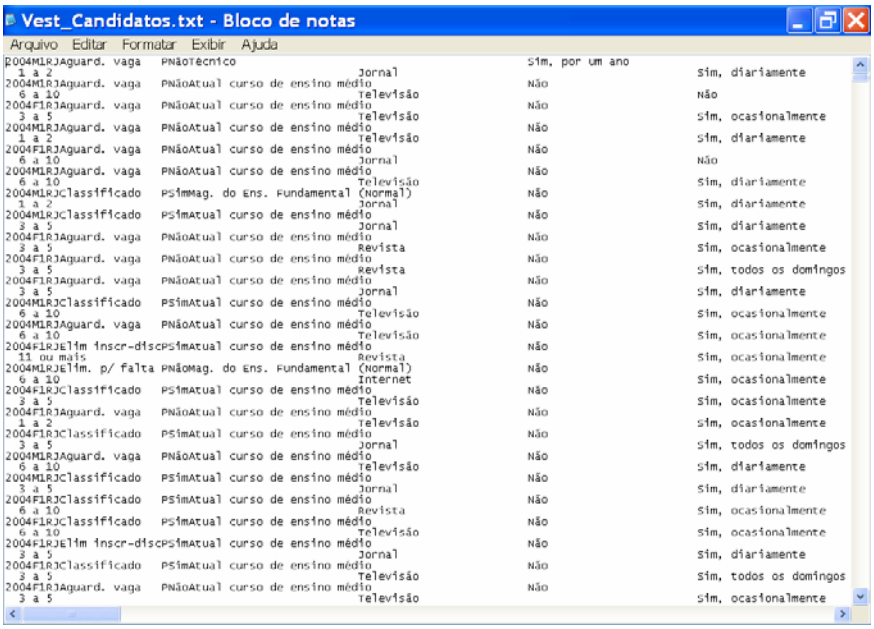



**Figura 14: Arquivo texto com os dados para mineração**

O Serviço de Mineração de Regras de Associação utiliza um arquivo de configuração de atributos, com a extensão `.atr`, para ler o arquivo texto contendo os campos selecionados do repositório de dados. Esse arquivo de atributos contém o nome dos campos, a posição da coluna onde ele começa e a quantidade de dígitos que ele possui. Cada linha contém um atributo. Os dois arquivos (atributos e dados) devem ser armazenados em uma pasta acessível pelo serviço de mineração. A Figura 15 apresenta um exemplo do arquivo de atributos, no formato texto.

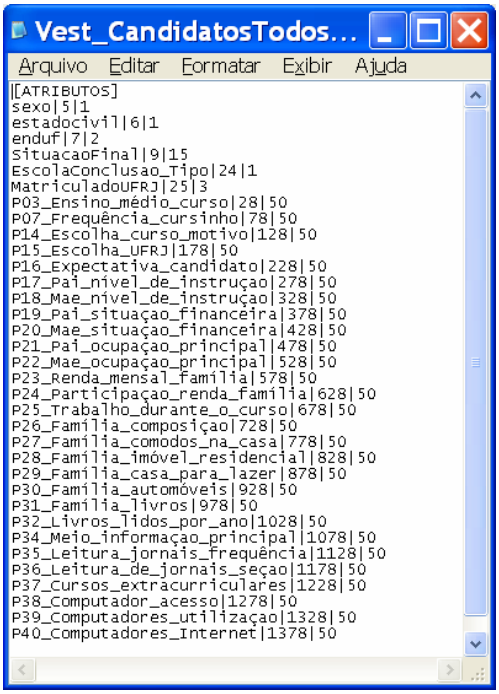
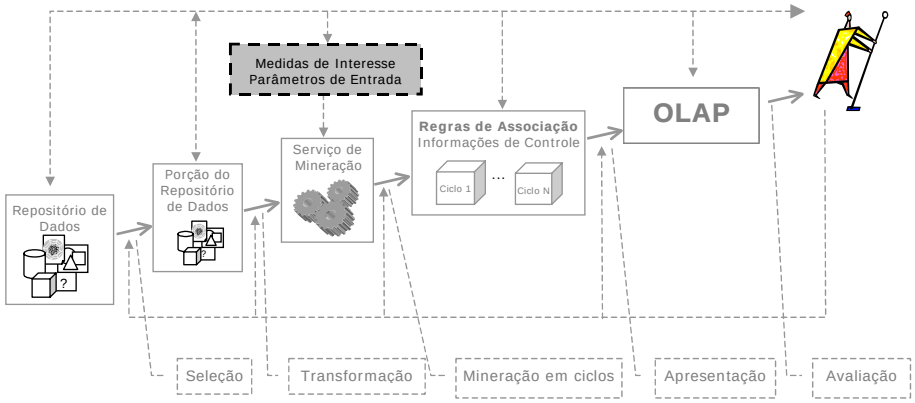



**Figura 15: Arquivo texto de atributos**

## 7.2 ETAPA 2 – DEFINIÇÃO DE MEDIDAS DE INTERESSE E PARÂMETROS DE ENTRADA




Na segunda etapa do processo analítico de exploração de regras de associação, as medidas de interesse e os parâmetros de entrada são definidos pelo usuário para alimentarem

o Serviço de Mineração de Regras de Associação. O usuário pode definir essas medidas utilizando conhecimento prévio sobre o negócio e consultando o repositório de dados ou a porção selecionada para a mineração.

As medidas de interesse e os parâmetros de entrada são informados ao Serviço de Mineração de Regras de Associação no seu módulo inicial. Esse módulo é aberto pelo ambiente de desenvolvimento, o *Microsoft Visual Studio*, utilizando um projeto *Visual Basic .Net*. O foco dessa dissertação não foi o desenvolvimento de uma interface amigável para utilização do serviço de mineração, por isso os parâmetros de entrada são definidos no próprio código do programa, conforme exibido na Figura 16.

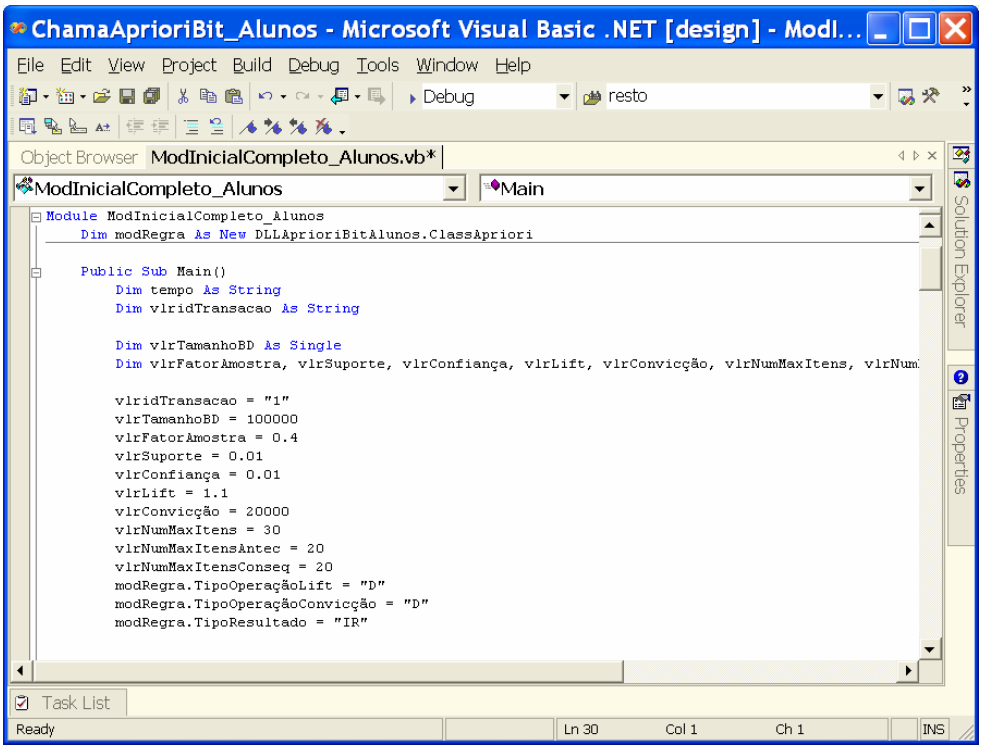
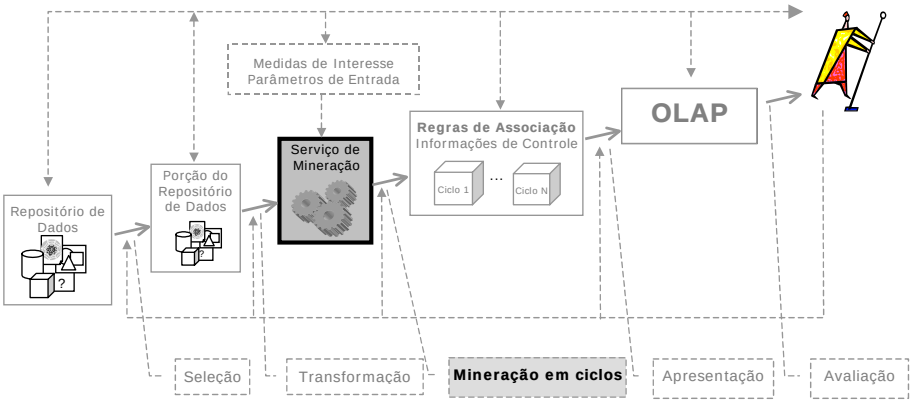



**Figura 16: Chamada do Serviço de Geração de Regras de Associação**

## 7.3 ETAPA 3 – EXECUÇÃO DO SERVIÇO DE MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO




O serviço de mineração (seção 5.1.2) utilizado neste exemplo de aplicação foi o Serviço de Geração de Regras de Associação de Lavôr (2003) que implementa, em *Microsoft Visual Basic*, o algoritmo *Apriori* para gerar as regras. Esse serviço foi adaptado à proposta para que a cada ciclo do algoritmo, as regras de associação candidatas e finais fossem geradas. Como foi dito na seção 6.1. por motivos de tempo e escopo, o serviço gera de uma só vez o resultado de todas as etapas. A proposta da abordagem é que apenas o resultado do primeiro ciclo seja gerado e que as outras etapas do processo (Apresentação e Avaliação) sejam executadas, oferecendo ao usuário pontos de interferência ao definir novas medidas de interesse e novos parâmetros de entrada, nova porção do repositório de dados, ou se o próximo ciclo será executado.

Após a seleção da porção do repositório de dados a ser minerada, a transformação dos dados dessa porção para o formato do serviço de geração de regras de associação e a definição das medidas de interesse e dos parâmetros de entrada, é executado, na terceira etapa do processo analítico de descoberta do conhecimento, o serviço de **mineração**, o qual gera as

amostras, os itens freqüentes de cada ciclo e as regras de associação. As barras de progresso, ilustradas na Figura 17, mostram o andamento da mineração.

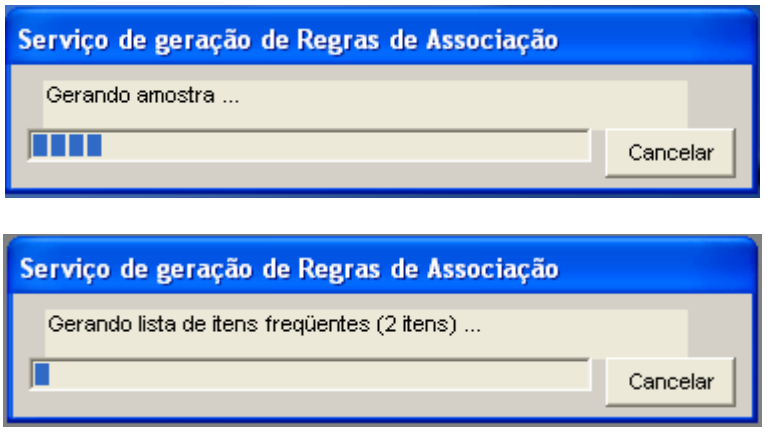



**Figura 17: Barras de Progresso do Serviço de Geração de Regras de Associação**

Ao final da execução, os arquivos listados na Figura 18, em formato texto, são gerados pelo serviço com o resultado da mineração. Esses arquivos serão utilizados para carregar o ambiente analítico com as informações de todo o processo de exploração das regras de associação, conforme apresentado na próxima etapa do processo.

| | |
|---|---|
| ItensFrq_Alunos.txt | Itens freqüentes |
| ItensFrq_Alunos_Candidatos_1.txt<br>ItensFrq_Alunos_Candidatos_2.txt<br>ItensFrq_Alunos_Candidatos_3.txt<br>ItensFrq_Alunos_Candidatos_4.txt<br>ItensFrq_Alunos_Candidatos_5.txt | Itens freqüentes candidatos de cada ciclo |
| Regras_Alunos_2.txt<br>Regras_Alunos_3.txt<br>Regras_Alunos_4.txt<br>Regras_Alunos_5.txt | Regras de Associação de cada ciclo |
| Regras_Candidatas_Alunos_2.txt<br>Regras_Candidatas_Alunos_3.txt<br>Regras_Candidatas_Alunos_4.txt<br>Regras_Candidatas_Alunos_5.txt | Regras de Associação candidatas de cada ciclo |
| Tempo.txt | Tempo de processamento |
| Transacao.txt | Dados da transação |

**Figura 18: Resultado da execução do Serviço de Geração de Regras de Associação**

Nesses arquivos, podemos encontrar: (i) o tempo gasto em todo o processamento, (ii) os dados da transação (tamanho do banco de dados, parâmetros e medidas de interesse

utilizados como entrada), (iii) os itens freqüentes e (iv) as regras candidatas e as selecionadas em cada ciclo do algoritmo *Apriori*.

A Figura 19 e a Figura 20, ilustram, respectivamente, exemplos dos arquivos texto dos itens freqüentes gerados pelo serviço e exemplos das regras de associação geradas.

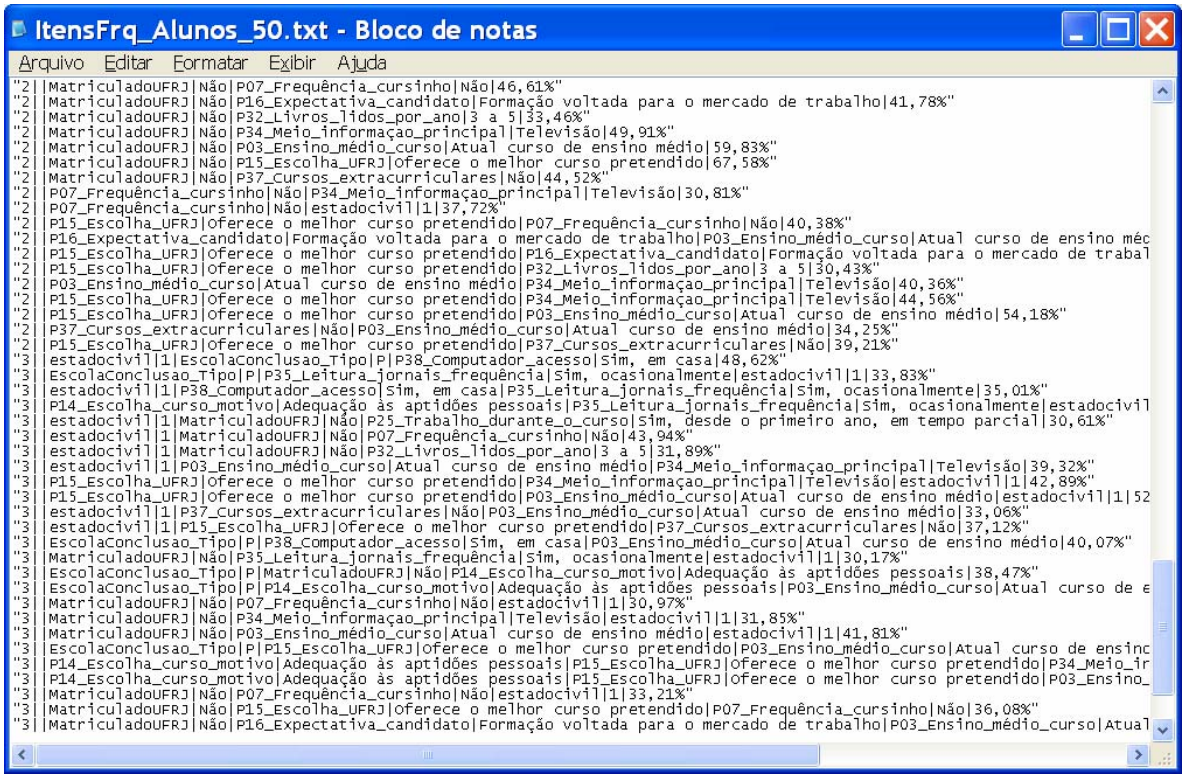



**Figura 19: Itens freqüentes gerados pelo serviço em um arquivo texto**

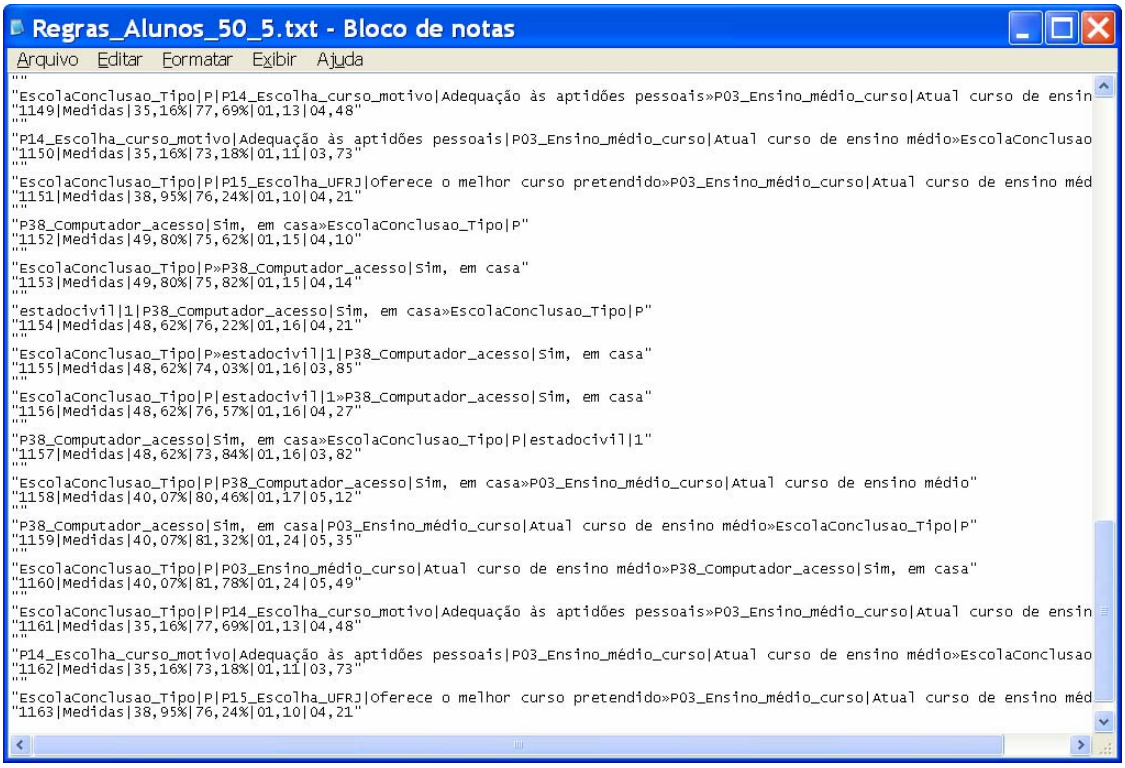
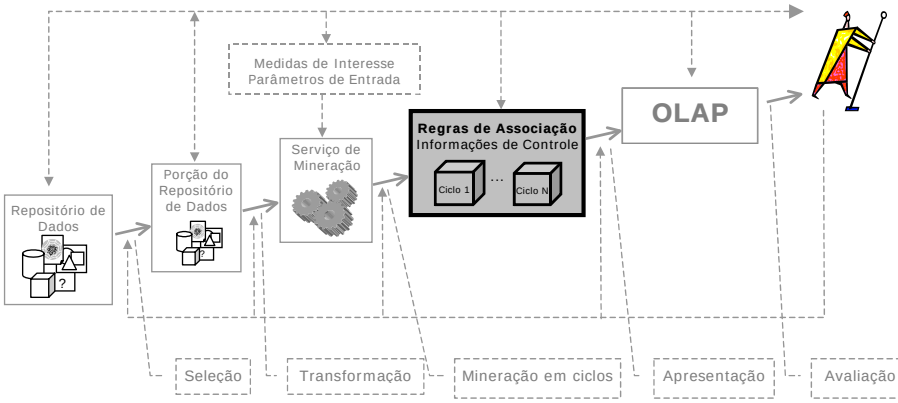



**Figura 20: Regras de Associação geradas pelo serviço em um arquivo texto**

## 7.4 ETAPA 4 – CARGA DO BANCO DE DADOS COM AS INFORMAÇÕES GERADAS PELO SERVIÇO




As regras, os itens freqüentes e as informações de controle, gerados pelo Serviço de Geração de Regras de Associação, são armazenados em arquivos em formato texto e, por isso, são difíceis de serem manipulados e analisados. Dessa forma, não atendem à primeira

característica de padrão interessante definido por Han e Kamber (2001): fácil de entender por humanos. Para atender a essa característica, é necessário que essas informações sejam levadas para um ambiente mais amigável e que permita a sua análise.

As informações dos arquivos gerados pelo serviço de mineração poderiam ser carregadas diretamente no banco de dados para serem analisadas pelo usuário. Porém, neste exemplo de aplicação, por motivos de tempo e escopo, o resultado é armazenado em arquivos texto e, depois, importado para o banco de dados. Dessa forma, a quarta etapa do processo analítico de exploração de regras de associação é a carga das informações geradas pelo serviço em um banco de dados, baseada no modelo de dados multidimensional definido na seção 6.2. As tabelas do modelo foram carregadas no banco de dados *Microsoft SQL Server 2005*. Para isso, rotinas de leitura, transformação e carga foram definidas utilizando o *Integration Services* do *Business Inteligence Development Kit* da *Microsoft* (Figura 21).

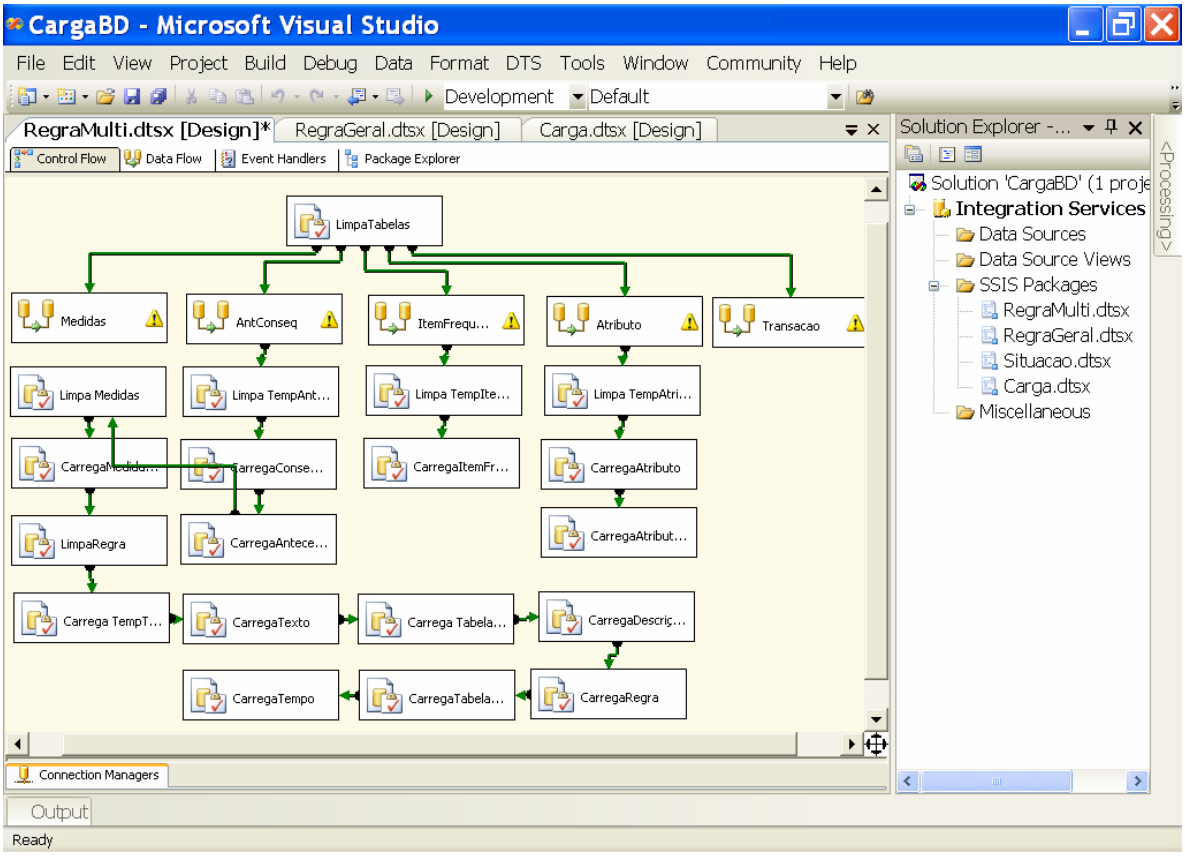



**Figura 21: Rotinas de leitura, transformação e carga do resultado da mineração**

A Figura 22 ilustra as tabelas do modelo de dados multidimensional definidas no banco de dados de regras de associação. Essas tabelas são utilizadas pela ferramenta OLAP, conforme descrito na quinta etapa do processo analítico de exploração de regras de associação.

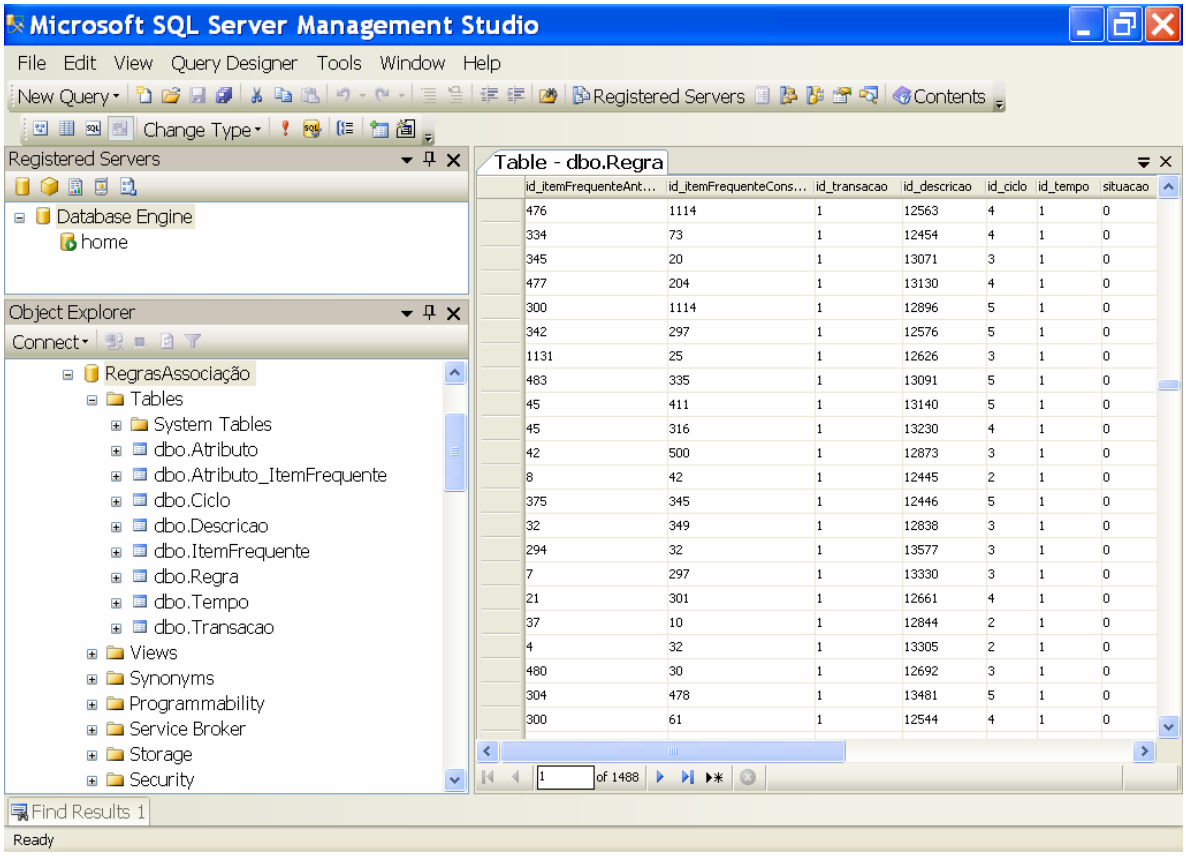
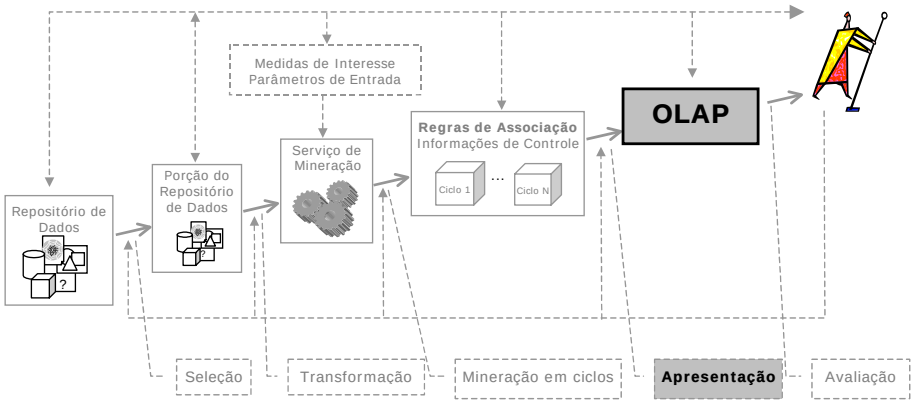



**Figura 22: Tabelas do modelo de dados multidimensional carregadas no banco de dados relacional**

## 7.5 ETAPA 5 – CARGA DO AMBIENTE ANALÍTICO




Na penúltima etapa do processo, o ambiente analítico é montado para que as informações geradas pelo serviço de mineração possam ser **apresentadas** em uma interface mais amigável e para que possam ser mais facilmente manipuladas pelo usuário.

O primeiro passo (Figura 23) da montagem do ambiente (*Microsoft Analysis Services*) é a definição da fonte de dados que, neste exemplo de aplicação, são as tabelas carregadas no banco de dados, conforme descrito na etapa anterior. O segundo passo (Figura 24) é a definição das dimensões, dos fatos e das métricas do cubo. E o terceiro passo (Figura 25) é o processamento (*deployment*) do cubo.

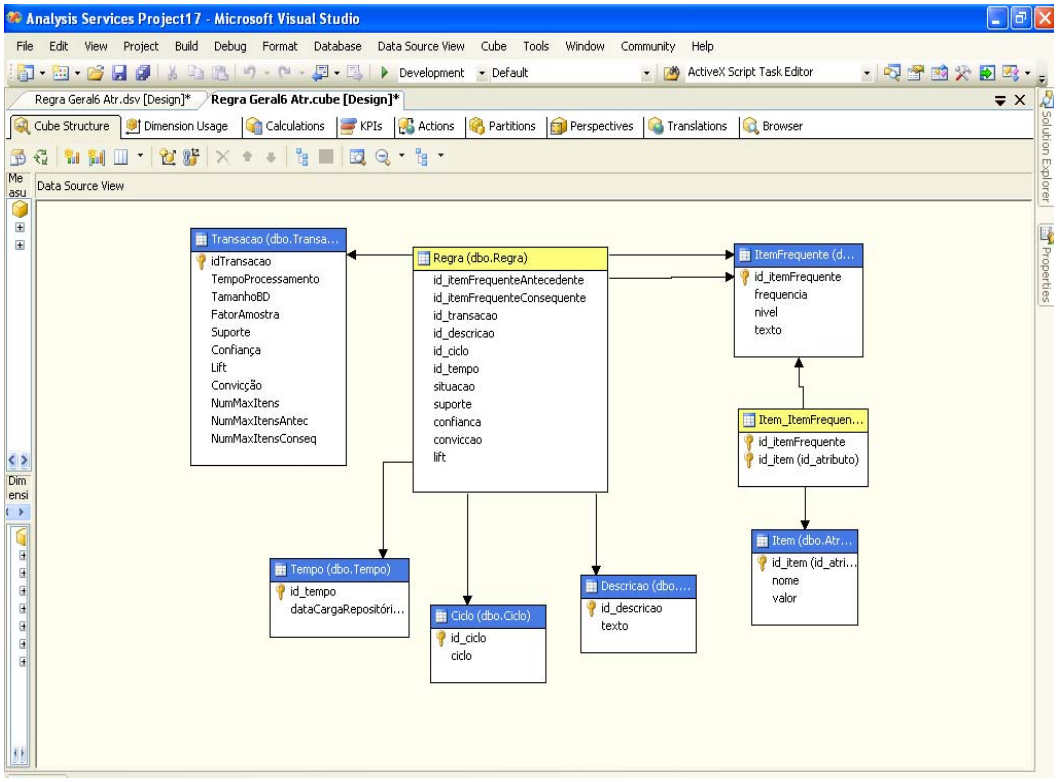



**Figura 23: Fonte de dados da ferramenta OLAP**

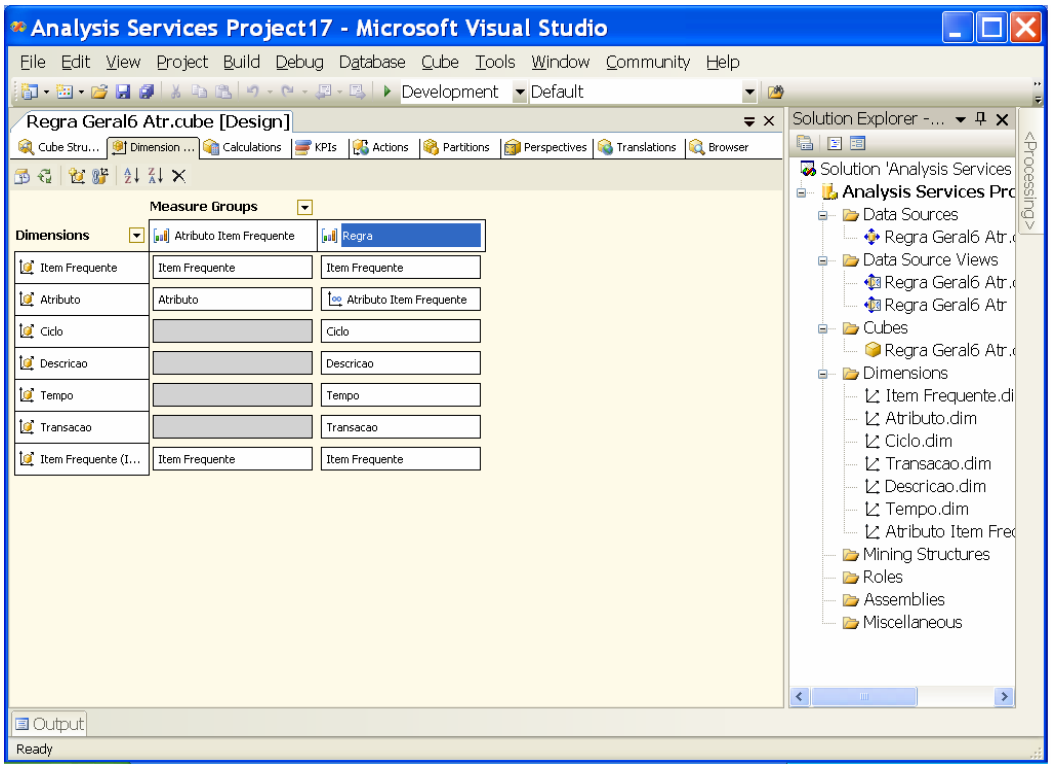



**Figura 24: Definição das dimensões, dos fatos e das métricas do cubo**

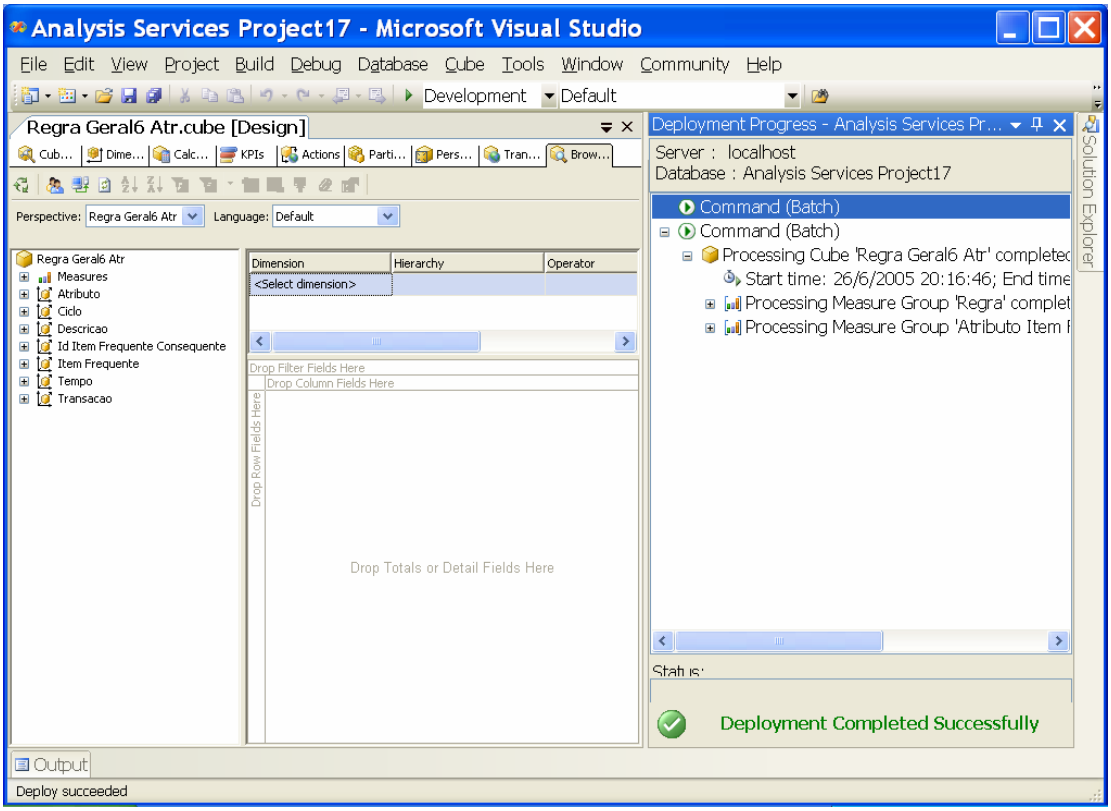
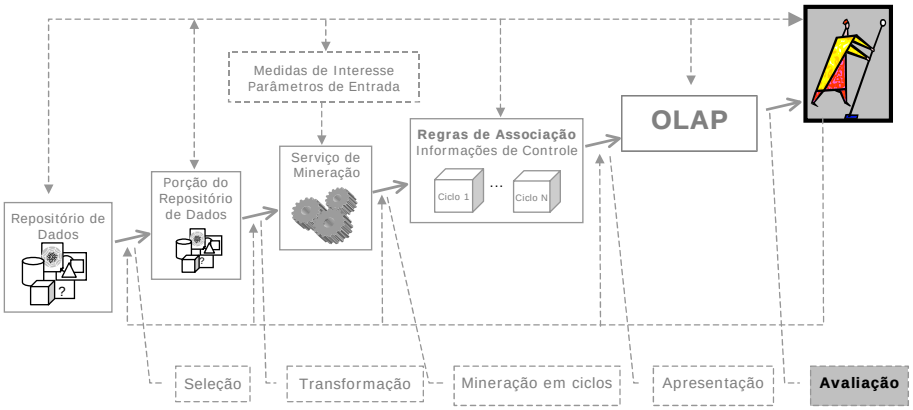



**Figura 25: Processamento (*deployment*) do cubo na ferramenta OLAP**

## 7.6 ETAPA 6 – ANÁLISE DOS DADOS DO PROCESSAMENTO E DAS REGRAS GERADAS




Na última etapa do processo analítico, os dados já foram selecionados, transformados, minerados e carregados no ambiente analítico para serem **analisados** pelo usuário. Observa-se que essa pode não ser a última etapa de fato, pois o usuário pode voltar às etapas anteriores e

definir novos parâmetros, em decorrência da análise realizada. Na figura do processo de descoberta do conhecimento, as setas pontilhadas e destacadas em negrito indicam todos os pontos em que o usuário pode voltar e executar o processo a partir daquele ponto. Neste exemplo de aplicação, cada uma das etapas é executada como aplicações isoladas, a abordagem propõe que todas essas etapas estejam integradas em um só aplicativo, oferecendo transparência para o usuário.

Observa-se também que o repositório de dados original, ou a porção selecionada, também são fontes de consulta que o usuário pode utilizar para enriquecer suas análises.

Com o ambiente analítico carregado, o usuário pode (i) visualizar o antecedente nas linhas, o conseqüente nas colunas e as quatro medidas de interesse na interseção entre as linhas e colunas (Figura 26); (ii) visualizar as regras descartadas e aproveitadas em cada ciclo da mineração (Figura 27); (iii) analisar as regras e suas quatro medidas de interesse (Figura 28); (iv) selecionar quais itens freqüentes devem aparecer no antecedente e conseqüente (Figura 29); (v) visualizar a quantidade de regras geradas por ciclo de uma determinada transação (Figura 30); (vi) visualizar a quantidade de regras que têm no seu antecedente um determinado item freqüente com todos os seus valores (Figura 31); (vii) visualizar a quantidade de regras por nível de item freqüente (Figura 32); (viii) visualizar a quantidade de regras geradas a cada ciclo, de acordo com as características de uma transação (Figura 33); (ix) quantidade de regras geradas em um processamento em uma carga do repositório (Figura 34); (x) suporte dos itens freqüentes no nível 1(um) (Figura 35); e (xi) aplicação de filtro para as regras de nível 1 (um) (Figura 36).

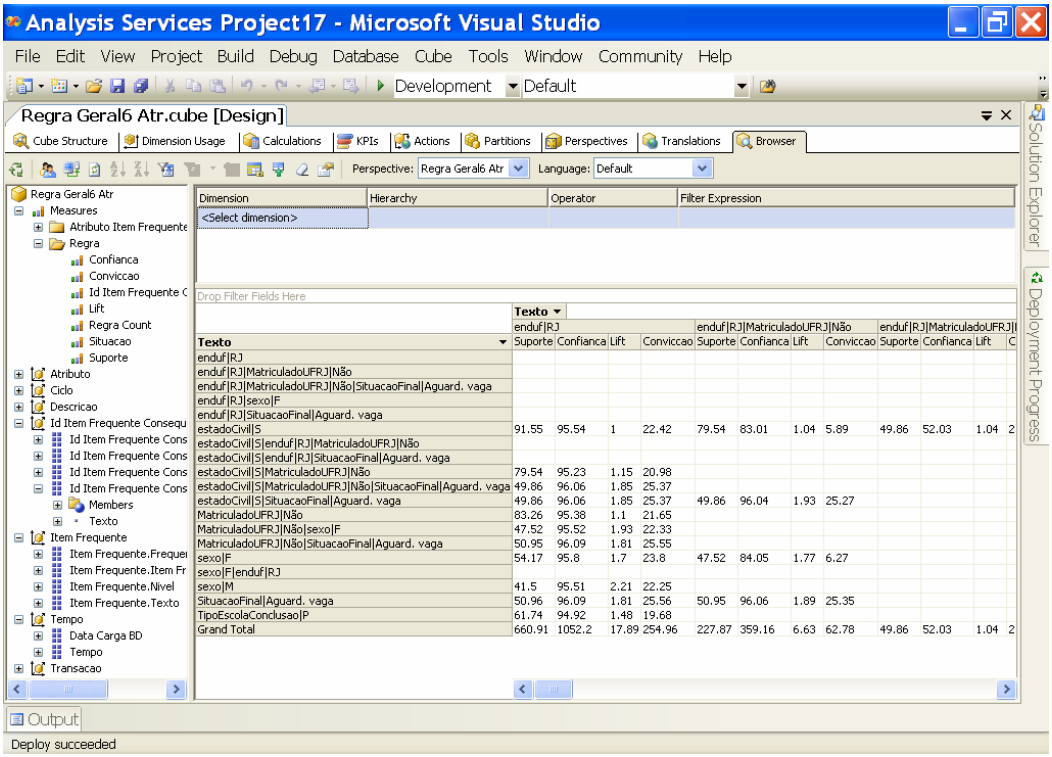



**Figura 26: Regras de Associação em um Ambiente Analítico**

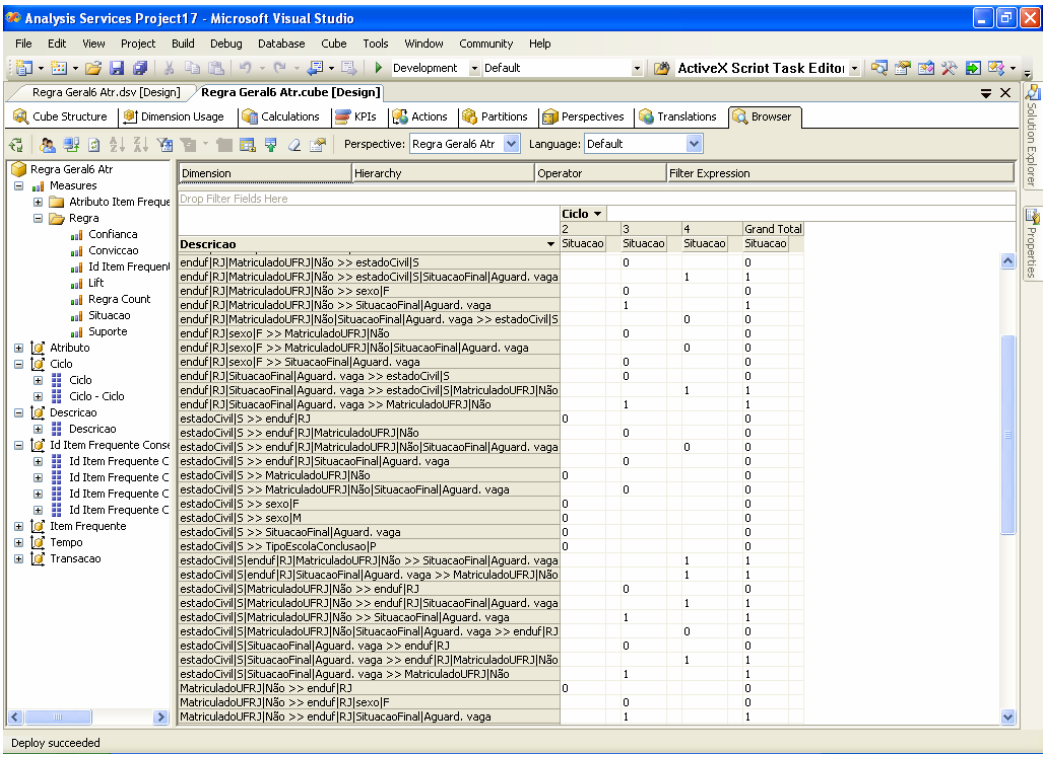



**Figura 27: Regras de Associação descartadas e aproveitadas em cada ciclo da mineração**

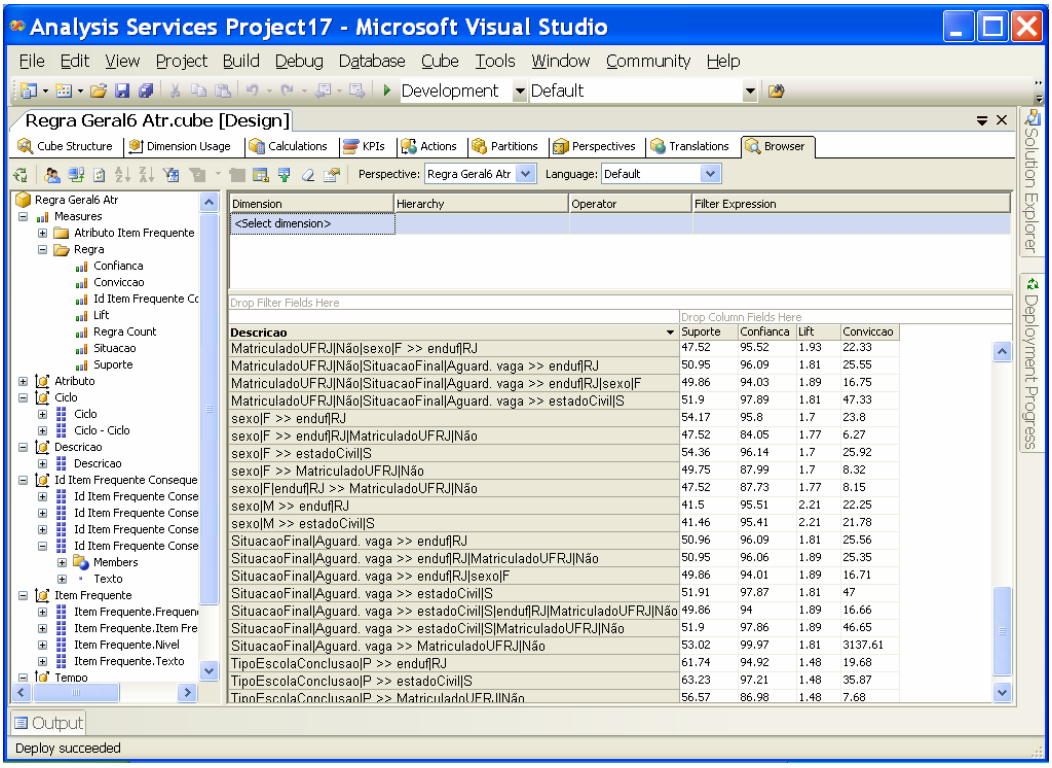



**Figura 28: Regas de Associação e suas medidas de interesse**

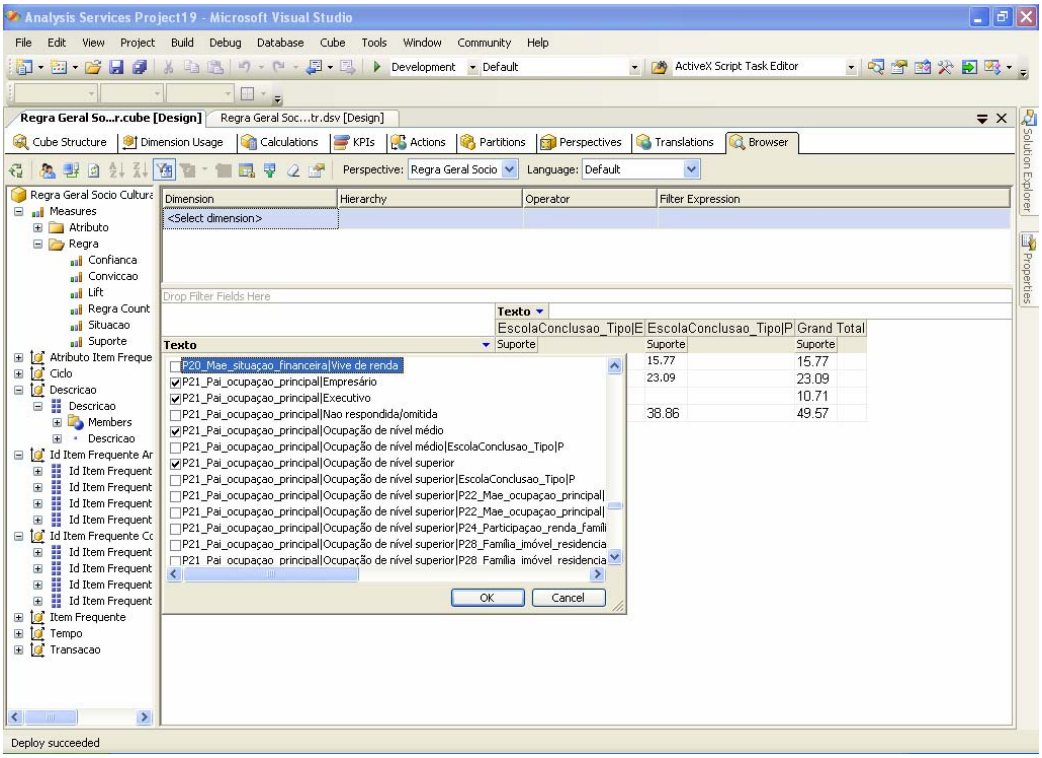



**Figura 29: Seleção do itens freqüentes do antecedente e do conseqüente**

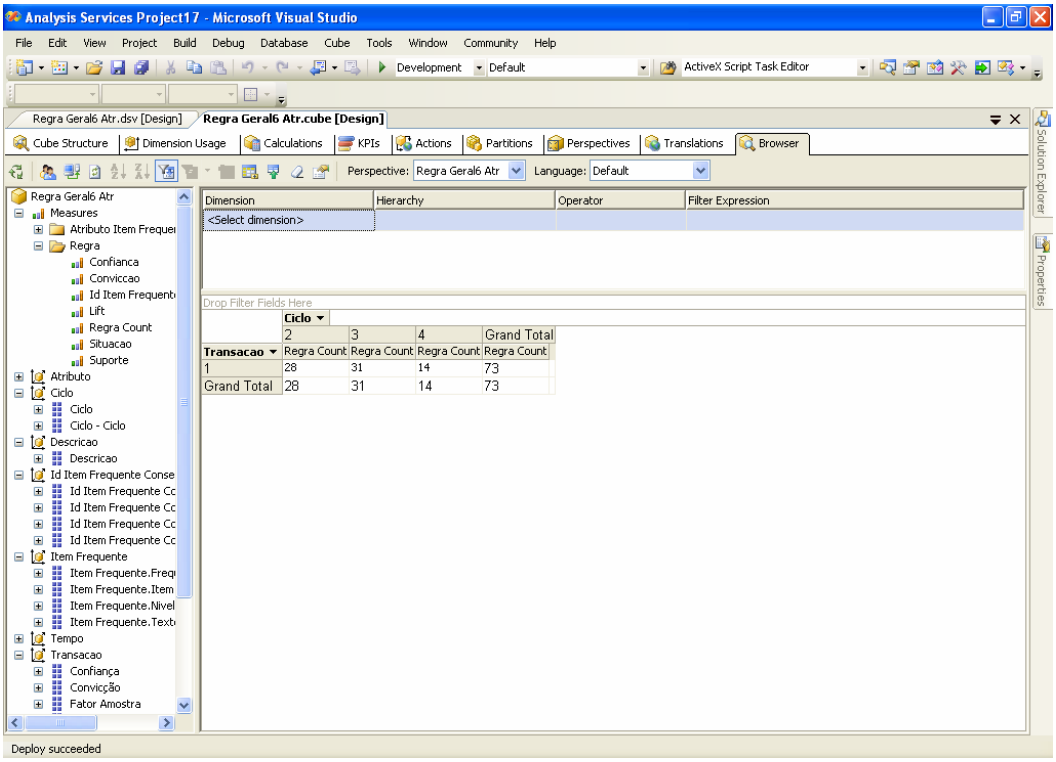



**Figura 30: Quantidade de regras geradas por ciclo de uma determinada transação**

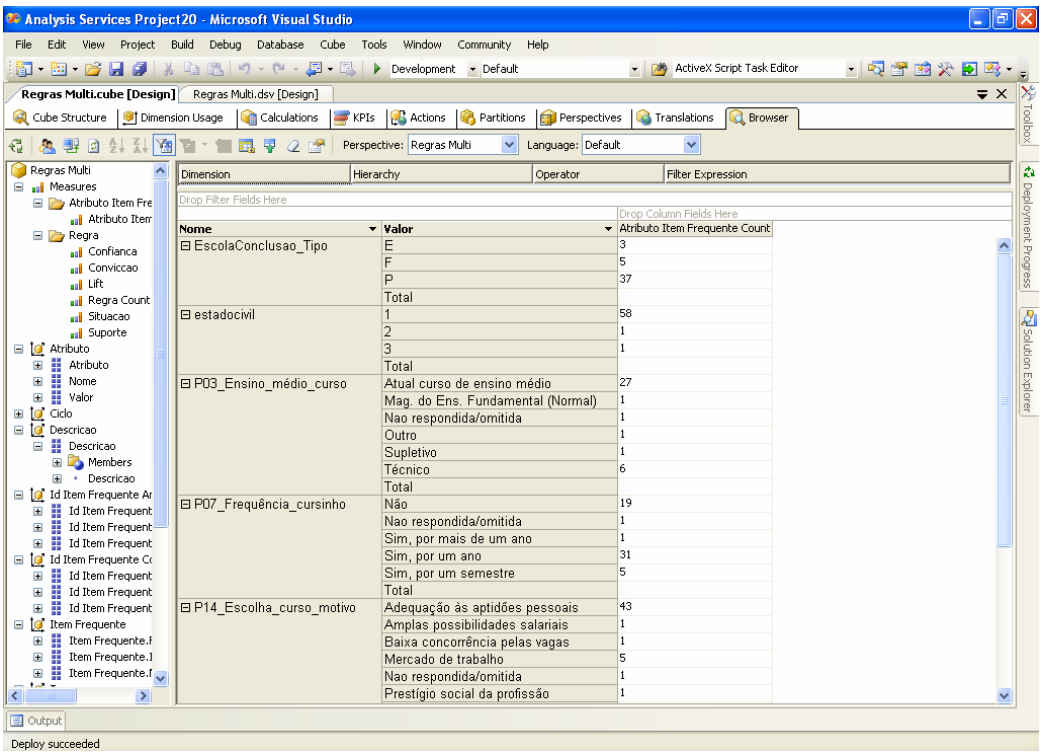



**Figura 31: Itens, seus valores e a quantidade de regras**

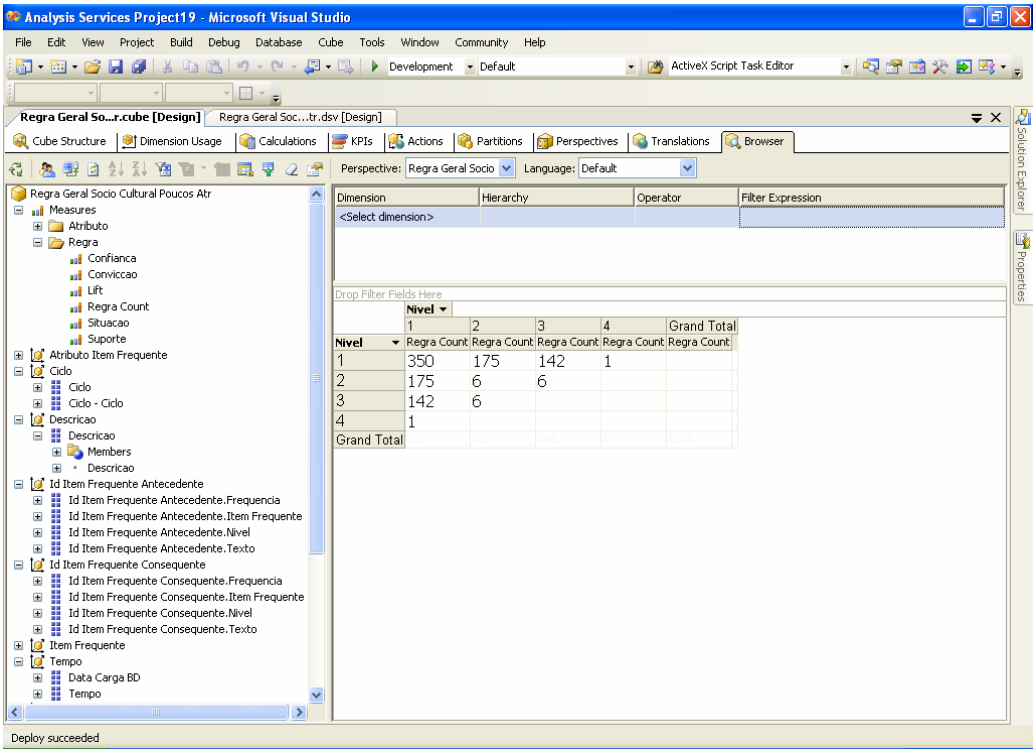



**Figura 32: Quantidade de regras de associação por nível de item freqüente**

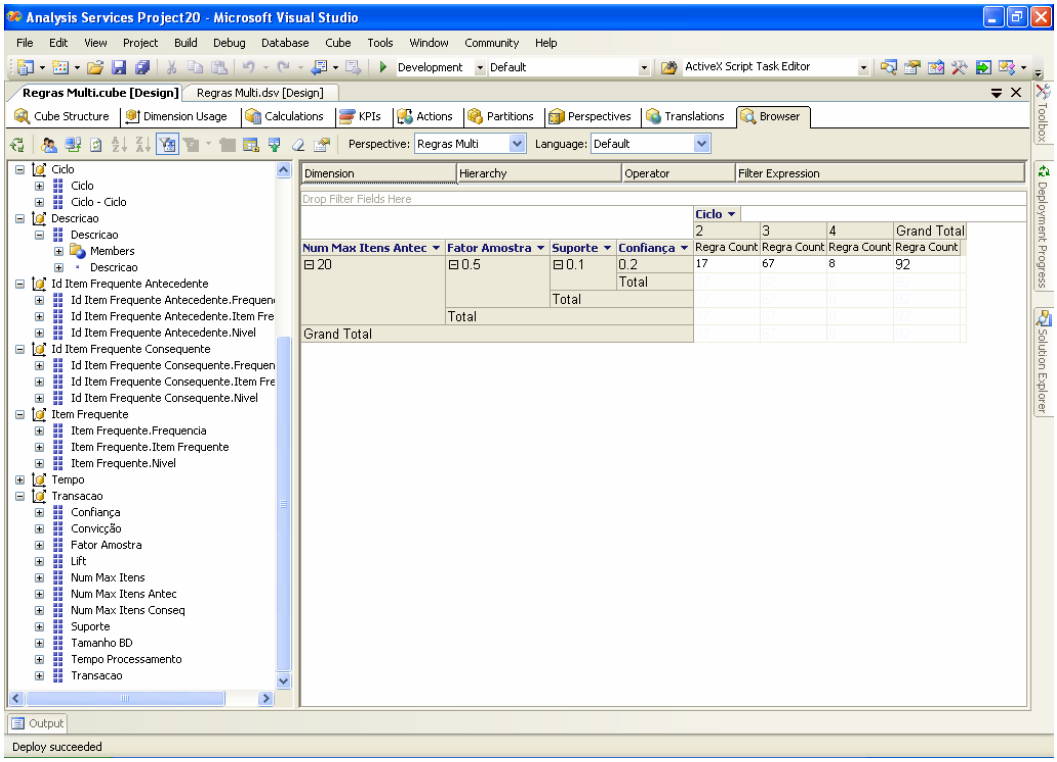



**Figura 33: Quantidade de regras geradas a cada ciclo, de acordo com as características de uma transação**

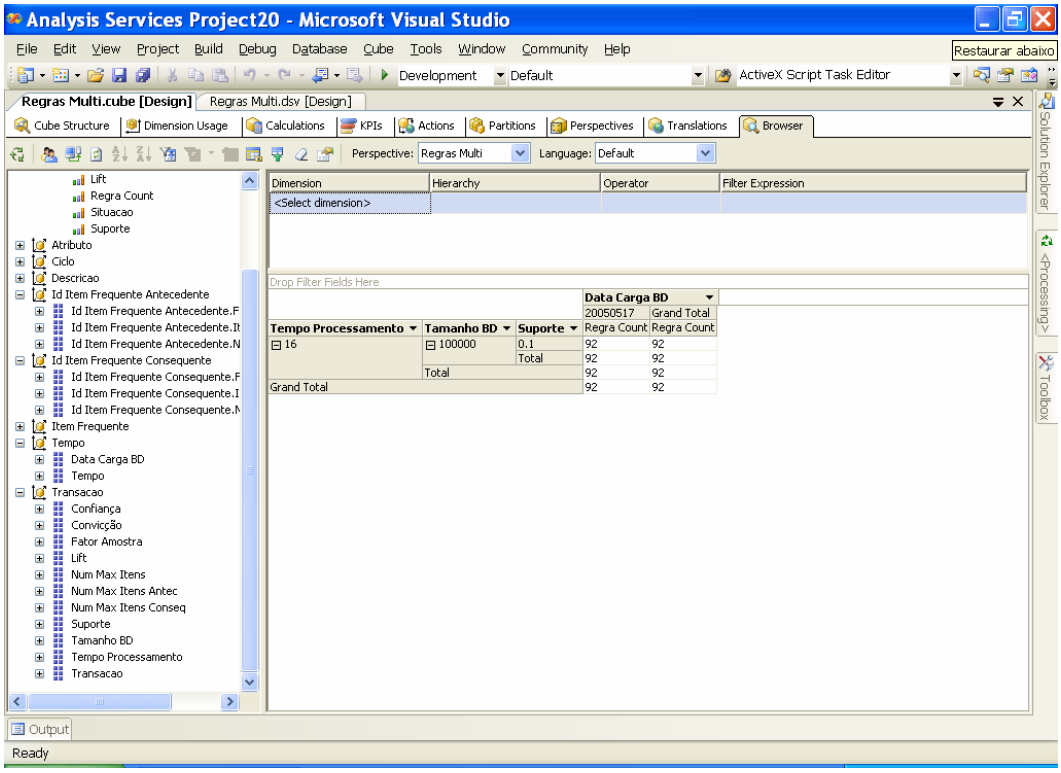



**Figura 34: Quantidade de regras geradas em um processamento em uma carga do repositório**

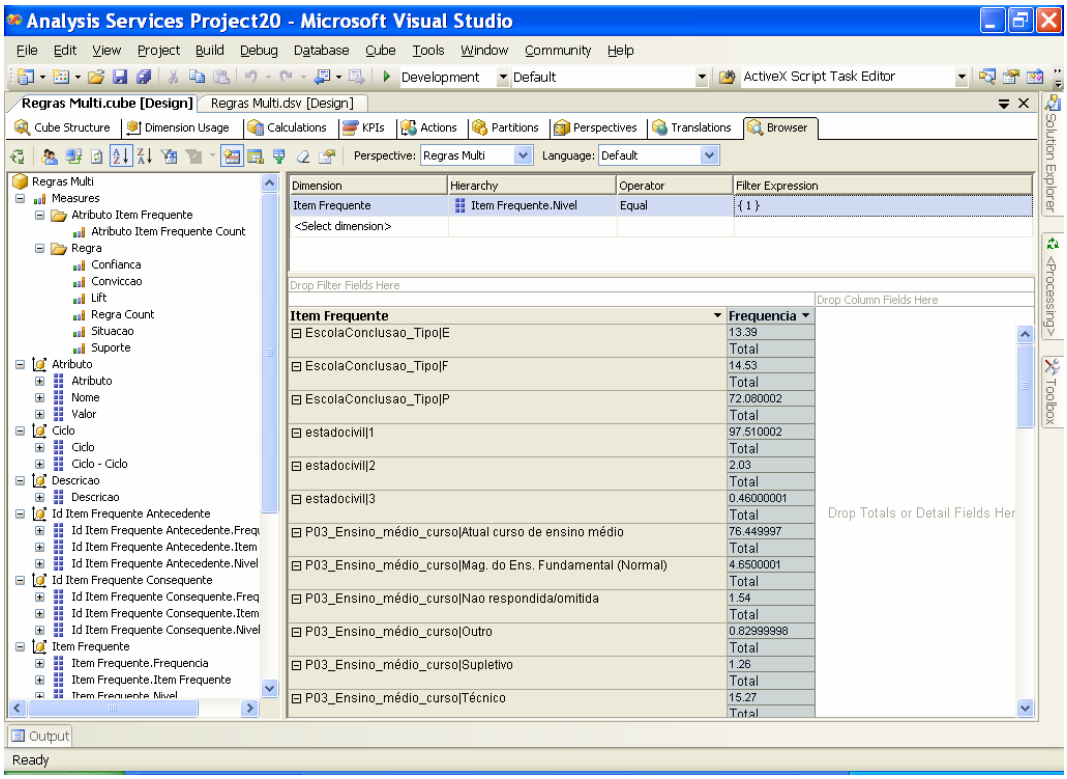



**Figura 35: Suporte dos itens freqüentes no nível 1(um)**

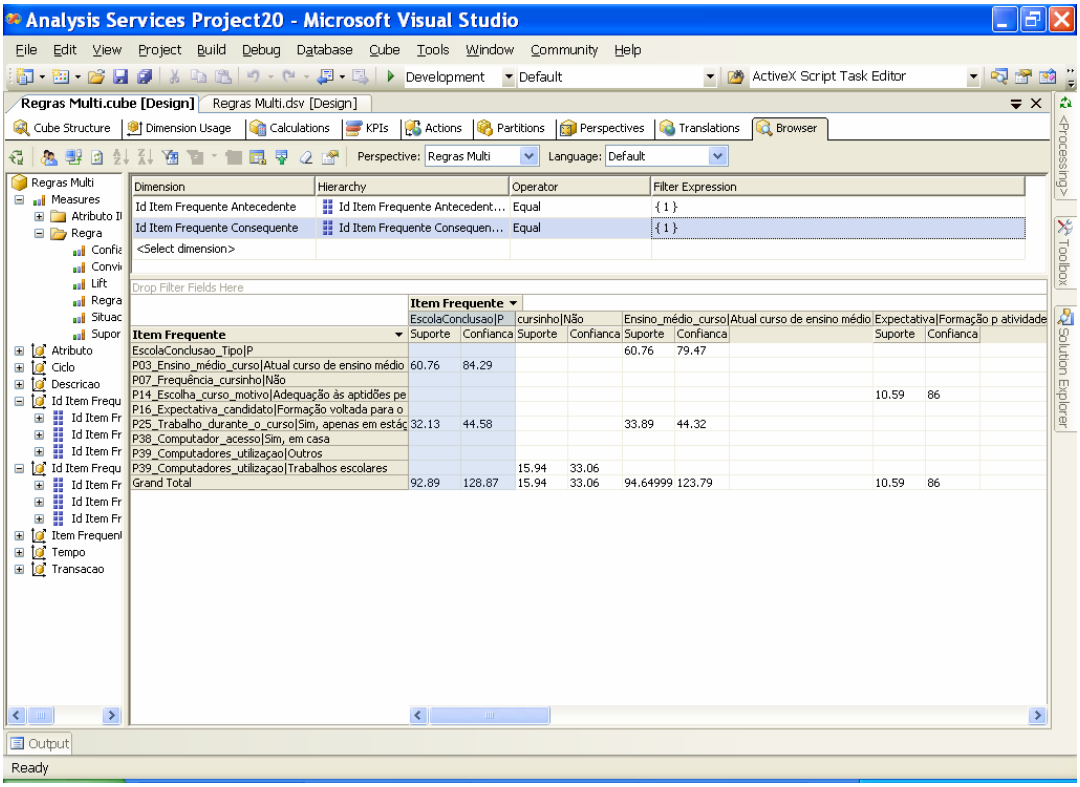



**Figura 36: Aplicação de filtro para as regras de nível 1 (um)**

# 8 CONCLUSÕES E TRABALHOS FUTUROS

Mineração de dados é uma tarefa de descoberta de padrões de dados interessantes em um conjunto de itens armazenados em um repositório. A descoberta de relacionamentos de associação é uma tarefa de mineração realizada por meio da aplicação de algoritmos em um conjunto de itens de dados. A mineração de regras de associação pode ajudar no processo de tomada de decisão ao revelar conhecimentos novos e úteis.

A tarefa de mineração de regras de associação, tipicamente, gera muitas regras, nem sempre interessantes e difíceis de serem manipuladas. A escolha dos padrões é feita pelos algoritmos de mineração com a ajuda de medidas - objetivas e subjetivas - definidas pelo usuário no início ou no final do processo. Esse procedimento pode gerar regras não pertinentes e deixar de fora regras interessantes.

Para contornar esse problema é necessária uma interação freqüente do usuário com o sistema, para realizar análises e direcionar a exploração do dado. As abordagens encontradas na literatura não oferecem um ambiente que permite a interferência do usuário no processo de exploração das regras. Dessa forma, regras interessantes podem ser excluídas do resultado final.

Este trabalho propõe um ambiente analítico de apoio ao processo de exploração de regras de associação que apresenta o caminho que está sendo percorrido para encontrar as associações. Com isso, permite que o usuário realize análises e interfira no processo, redirecionando-o para regras mais interessantes. A exploração de regras de associação em um ambiente analítico permite estudo, análise e avaliação de todo o caminho de mineração percorrido e do conhecimento adquirido.

O ambiente analítico proposto visa apoiar o analista de negócios no processo de tomada de decisão, oferecendo um ambiente fácil de usar e entender. Entretanto, é necessário

que o usuário tenha conhecimento técnico e específico do processo de mineração de regras de associação, e habilidades para operar uma ferramenta OLAP.

As principais contribuições desta pesquisa são as definições da arquitetura do ambiente e do processo analítico para exploração de regras de associação, e o modelo de dados multidimensional de regras de associação. O ambiente analítico permite avaliar o que é útil e a interferência do usuário no processo permite selecionar o que interessa ou não, possibilitando, então, a redução da quantidade de regras geradas. A solução apresentada oferece um ambiente analítico de alto nível, com facilidades de interação, que permite a integração de dados conhecidos com um mecanismo de mineração, possibilitando o direcionamento da exploração de padrões.

A análise das regras geradas a cada ciclo do algoritmo *Apriori* foi inserida como mais um passo no processo de descoberta do conhecimento. Todos os passos desse processo foram implementados, mostrando a viabilidade prática da proposta. Porém se faz necessário a implementação de uma ferramenta que integre todas as etapas do processo e que o desempenho seja tratado. O ambiente analítico proposto não é dependente de uma plataforma tecnológica específica, assim oferece flexibilidade na sua aplicação.

Como trabalhos futuros, sugerimos:

- Uma avaliação sistemática, com a aplicação de uma metodologia, para verificar se a quantidade de regras geradas foi reduzida, em comparação às outras abordagens, e se as mais interessantes estavam no resultado final. Essa avaliação envolve questões humanas, tais como: o conhecimento do usuário sobre a área de negócio e sua habilidade em operar ferramentas de mineração e de análise de dados;
- O tratamento do desempenho do processo de exploração;
- A implementação de uma ferramenta que integre todas as etapas do processo;

- A definição de métodos e técnicas que permitam confrontar os padrões descobertos com os dados originais; e
- A aplicação de técnicas de sistemas especialistas que habilitem o ambiente a aprender com as decisões tomadas pelo usuário ao longo do tempo e a utilizar esse conhecimento para as novas minerações, diminuindo, assim, as necessidades de intervenção do usuário.

# REFERÊNCIAS

AGGARWAL, C. C. e YU, P. S. **A new framework for itemset generation**. In: Symposium on Principles of Database Systems, 1998, Seattle, WA, USA, p. 18-24

AGRAWAL, R.; AGGARWAL, C. C. e PRASAD, V. V. V. **A tree projection algorithm for generation of frequent item sets**, 2001, Journal of Parallel and Distributed Computing, v.61, p. 350-371

AGRAWAL, R.; IMIELINSKI, T. e SWAMI, A. **Mining Association Rules between Sets of Items in Large Databases**. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1993, Washington, D.C, USA, p. 207-216

AGRAWAL, R. e SRIKANT, R. **Fast algorithms for mining association rules**. In: Proceedings of the International Conference on Very Large Data Bases, 1994, Santiago, Chile, p. 487-499

BRIN, S.; MOTWANI, R. e ULLMAN, J. D. **Dynamic Itemset Counting and Implication Rules for Market Basket Data**. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1997, Tucson, Arizona, USA, p. 255-264

BRUSSO, M. J. **Access Miner: Uma Proposta para a Extração de Regras de Associação Aplicada a Mineração do Uso da Web.** 2000. 96f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2000

BRUZZESE, D. e BUONO, P. **Combining visual techniques for Association Rules exploration**. In: Proceedings of the Working Conference on Advanced Visual Interfaces, 2004, Gallipoli, Italy, p. 381-384

BURDICK, D.; CALIMLIM, M. e GEHRKE, J. **MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases**. In: International Conference on Data Engineering, 2001, Washington, DC, USA, p. 443-452

CAMPOS, M. L. M. **OLAP e Ferramentas de Acesso em Ambientes de DW**, Nota de Aula. Tópicos Especiais em Banco de Dados II. Departamento de Ciência da Computação - Instituto de Matemática - Universidade Federal do Rio de Janeiro, 2004. Acessado em: 21-11-2004. Disponível em: http://dataware.nce.ufrj.br:8080/dataware/Cursos/Graduacao/tebd2/modulos/dataware/fisico/apresentacoes/datawarehouse/Mod32004.ppt

DHAR, V. e TUZHILIN, A. **Abstract-driven pattern discovery in databases**. In: IEEE Educational Activities Department, 1993, Piscataway, NJ, USA, v. 5, p. 926-938

DONG, G. e LI, J. **Interestingness of Discovered Association Rules in Terms of Neighborhood-Based Unexpectedness Source**. In: Proceedings of the Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining, 1998, London, UK, p. 72-86

FUKUDA, T.; MORIMOTO, Y; MORISHITA, S. e TOKUYAMA, T. **Data Mining using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization**. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1996, Montreal, Quebec, Canada, p. 13-23

GOUDA, K. e ZAKI, M. J. **Efficiently Mining Maximal Frequent Itemsets**. In: IEEE International Conference on Data Mining, 2001, Washington, DC, USA, p. 163-170

GRAHNE, G. e ZHU, J. **High Performance Mining of Maximal Frequent Itemsets**. In: Proceeding of the International Workshop on High Performance Data Mining, 2003, Washington, DC, USA, p. 10-16

HAN, J. **OLAP Mining: An Integration of OLAP with Data Mining**. In: Proceedings IFIP Conference on Data Semantics, 1997, Kansas City, Missouri, USA, p. 1-11

HAN, J. e KAMBER, M. **Data Mining: Concepts and Techniques**, San Francisco:Morgan Kaufmann Publishers, 2001, 550 f.

HAN, J.; PEI, J. e YIN, Y. **Mining Frequent Patterns without Candidate Generation**. In: Proceedings ACM SIGMOD International Conference on Management of Data, 2000, Dallas, Texas, USA, p. 1-12

HIDBER, C. **Online association rule mining**. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1999, Philadelphia, PA, USA, p. 145-156

HOUTSMA, M. e SWAMI, A. **Set-Oriented Mining of Association Rules**. In: IBM Almaden Research Center, 1993, San Jose, California, USA, p. 25-34

INMON, W. H. **Como construir o data warehouse**, Rio de Janeiro, RJ:Campus, 1997, 404 f.

KAMBER, M.; HAN, J. e CHIANG, J. Y. **Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes**. In: International Conference on Knowledge Discovery and Data Mining, 1997, Newport Beach, California, USA, p. 207-210

KIMBALL, R. **Causal (Not Casual) Dimensions**, 1996. Acessado em: 7-8-2005. Disponível em: http://www.dbmsmag.com/9611d05.html

KIMBALL, R. **The Soul of the Data Warehouse, Part Two: Drilling Across**, 2003. Acessado em: 7-8-2005. Disponível em: http://www.intelligententerprise.com/030405/606warehouse1_1.shtml

KIMBALL, R.; REEVES, L.; ROSS, M. e THORNTHWAITE, W. **The Data Warehouse Lifecycle Toolkit**,Wiley, 1998, 800 f.

KLEMETTINEN, M.; MANNILA, H.; RONKAINEN, P.; TOIVONEN, H. e VERKAMO, A. I. **Finding Interesting Rules from Large Sets of Discovered Association Rules**. In: International Conference on Information and Knowledge Management, 1994, Gaithersburg, Maryland, USA, p. 401-407

LAVÔR, R. M. P. **Implementação de serviços relacionados à mineração de regras de associação.** 2003. 152f. Dissertação (Mestrado em Informática) - Instituto de Matemática/Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, 2003

LIU, B.; HSU, W.; MUN, L. e LEE, H. **Finding Interesting Patterns Using User Expectations**. In: IEEE Transactions on Knowledge and Data Engineering, 1999, v. 11, p. 817-832

LUCCHESE, C.; ORLANDO, S. e PEREGO, R. **DCI Closed: A Fast and Memory Efficient Algorithm to Mine Frequent Closed Itemsets**. In: Workshop on Frequent Itemset Mining Implementations, 2004, Brighton, UK

MONTEIRO, R. S.; ZIMBRÃO, G. e SOUZA, J. M. **An Analytical Approach for Handling Association Rule Mining Results**. In: Australiasian Data Mining WorkShop, 2003, Canberra, Australia

MORISHITA, S. **On Classification and Regression**. In: International Conference on Discovery Science, 1998, Fukuoka, Japan, p. 40-57

OMIECINSKI, E. R. **Alternative interest measures for mining associations in databases**. In: IEEE Transactions on Knowledge and Data Engineering, 2003, v. 15, p. 57-69

PADMANABHAN, B. e TUZHILIN, A. **Unexpectedness as a Measure of Interestingness in Knowledge Discovery**. In: Decision Support Systems, 1999, Amsterdam, v. 27, p. 303-318

PIATETSKY-SHAPIRO, G. **Discovery, analysis, and presentation of strong rules**. In: Knowledge Discovery in Databases, 1991, p. 229-248

PIATETSKY-SHAPIRO, G. e MATHEUS, C. J. **The interestingness of deviations**. In: Workshop on Knowledge Discovery in Databases, 1994, Seattle, WA, USA, p. 25-36

QUINLAN, J. R. **Program for Machine Learning**.1992, San Mateo, California, USA

REINSCHMIDT, J.; GOTTSCHALK, H.; KIM, H. e ZWIETERING, D. **Intelligent Miner for Data: Enhance Your Business Intelligence**, San Jose, California, USA:IBM Corporation, International Technical Support Organization, 1999, 228 f.

SENO, M. e KARYPIS, G. **LPMiner: An Algorithm for Finding Frequent Itemsets Using Length-Decreasing Support Constraint**. In: IEEE International Conference on Data Mining, 2001, San Jose, California, USA, p. 505-512

SIEBES, A. e FEELDERS, Ad **Association Rules**, 2003. Acessado em: 20-11-2004. Disponível em: http://www.cs.uu.nl/docs/vakken/dm/assoc-complete.pdf

SILBERSCHATZ, A. e TUZHILIN, A. **On Subjective Measures of Interestingness in Knowledge Discovery**. In: International Conference on Knowledge Discovery and Data Mining, 1995, Montreal, Quebec, Canada, p. 275-281