Mineração da Base de Dados de um Processo Seletivo Universitário

ARACELE GARCIA DE OLIVEIRA
DENISE FERREIRA GARCIA

UNIFOR/MG – Centro Universitário de Formiga
ICSAE - Instituto de Ciências Sociais Aplicadas e Exatas
Curso de Ciência da Computação
Av. Dr. Arnaldo de Senna, 328 - Água Vermelha
CEP: 35570-000 - Formiga - MG - Brasil
(aracele,denise) @uniformg.edu.br

Resumo. O presente artigo descreve testes realizados e os resultados obtidos com a aplicação das etapas iniciais do Processo de Descoberta de Conhecimento (KDD) e a técnica de Regras de Associação da etapa de Data Mining. Elas foram utilizadas sobre os dados relacionados ao Questionário Sócio-Econômico-Cultural aplicado durante o Processo Seletivo do Centro Universitário de Formiga no ano de 2004. O objetivo maior do projeto aqui relatado é encontrar informações úteis que possam se transformar em conhecimento estratégico e que possam estar escondidas dentro da base de dados do Processo Seletivo do UNIFOR. Objetiva, também, demonstrar a importância dos cursos de computação universitários frente às aplicações dessas tecnologias como fator de apoio à competitividade dentro do mercado econômico atual.

Palavras-Chave: KDD, Data Mart, Data Mining, Descoberta de Regras de Associação, Processo Seletivo.

1 Introdução

Informação e Tecnologia caminham juntas desde que os computadores se tornaram equipamentos comerciais, na década de 60. A partir desse momento, imensos volumes de dados têm sido sistematicamente coletados e armazenados nas Bases de Dados das Empresas.

Essas Bases são uma importante fonte de informação, porém, muitas vezes, não são exploradas dadas as dificuldades inerentes a este grande volume, ultrapassando assim a habilidade técnica e a capacidade humana em sua interpretação [Carvalho(1999)]. Esses dados passam a possuir a devida importância dentro da empresa quando são tratados e explorados de forma adequada, utilizando para isso as diversas tecnologias que surgem ou que são aprimoradas a cada dia como KDD (Knowledge Discovery in Databases), OLAP (On-Line Analytical Processing), BI (Business Inteligence) e Data Mining, valiosas no apoio à tomada de decisão.

Também nas Instituições de Ensino, onde o Capital Intelectual geralmente está inserido, pode-se utilizar todo o conhecimento descoberto para melhorar a qualidade dos serviços prestados e favorecer ainda mais os recursos para que mais conhecimentos aliados à sabedoria e à experiência sejam disseminados não só no meio acadêmico como para toda a comunidade. Esse investimento é importante, considerando também a forte competição dentro do Mercado Educacional, onde o número de Instituições e cursos criados a cada ano vem se tornando cada vez maior.

Frente a essa realidade e às constantes mudanças por que tem passado, o Centro Universitário de Formiga faz da busca pela qualidade da Informação uma constante e

o tratamento despendido aos dados tem sido cada vez mais aperfeiçoado.

Diante disso, verificou-se a necessidade de explorar a Base de Dados do processo seletivo, aplicando sobre esta algumas etapas do Processo de Descoberta de Conhecimento – KDD a fim de identificar informações que poderiam estar "escondidas".

2 O Processo de Descoberta de Conhecimento em Bases de dados

A definição de Descoberta de Conhecimento em Bases de Dados, KDD (Knowledge Discovery Databases), foi definida por Fayyad como sendo: "... o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, embutidos nos dados" [Fayyad (1996)].

A Extração de Conhecimento é uma área multidisciplinar que incorpora técnicas utilizadas em diversas áreas como Banco de Dados, Inteligência Artificial e Estatística. Desse modo, as técnicas utilizadas em KDD não devem ser vistas como substitutas de outras formas de análise, como o OLAP, mas, sim, como práticas para melhorar os resultados das explorações feitas com as ferramentas atualmente usadas [Zambon & Meirelles (2001)].

2.1 Etapas do Processo de Descoberta de Conhecimento em Bases de Dados

O processo de KDD é caracterizado como sendo um processo interativo e iterativo, composto por várias etapas interligadas [Fayyad (1996)]. Essas etapas vão deste a definição do domínio, seleção, preparação e transformação dos dados, até a etapa de Data Mining, onde padrões podem ser "descobertos" e analisados para tornarem-se conhecimento útil.

A primeira etapa dentro do processo é a compreensão e definição de um domínio. Logo após, é necessário selecionar, dentro deste domínio, os dados nos quais o "descobrimento" será realizado. Estes dados devem ser limpos e transformados. Essa limpeza inclui a remoção de ruídos, a adequação de valores que estejam fora do contexto, a seleção e o resumo das variáveis a serem utilizadas.

É preciso, ainda, definir a técnica e o algoritmo de Data Mining a serem utilizados. Assim, os dados selecionados do domínio devem ser então transformados de acordo com as características da técnica e do algoritmo. Neste ponto, os dados já podem ser submetidos ao processo de Mineração propriamente dito.

A partir daí, com o resultado gerado, pode-se analisar o conhecimento descoberto. E caso os resultados não sejam satisfatórios, várias etapas do processo podem ser realizadas novamente.

Várias tecnologias que trabalham com manipulação de repositórios de dados podem ser utilizadas como apoio ao Processo de Descoberta de Conhecimento, como Data Warehouses e Data Marts.

Data Marts são Data Warehouses departamentais e estão sendo amplamente utilizados como repositórios de dados para a etapa de mineração [Italiano & Esteves (2004)]. Os dados neles contidos já foram selecionados, de acordo com um domínio específico, além de limpos, corrigidos e adequados. Estes dados se encontram prontos para serem submetidos ao processo de mineração em busca de conhecimento.

O modelo estrutural para representar um Data Mart é o Modelo Multidimensional, sendo o Estrela o tipo de modelo mais utilizado. No modelo Estrela, existe uma tabela com múltiplas chaves denominada tabela Fato e um conjunto de outras tabelas denominadas Dimensões. A transferência e o tratamento dos dados entre a base operacional e o Data Mart podem ser realizados com o apoio do DTS.

O DTS (Data Transformation Services) é uma ferramenta bastante útil que facilita este processo de Extração de Conhecimento pois auxilia no processo de importação, transformação e carregamento dos dados, denominado ETL (Extraction, Trasformation and Load) [Pichiliani (2004)]. Através da ferramenta, selecionamos as Fontes de Dados que podem fazer parte do Data Mart. Este dados são limpos, transformados e exportados para o Repositório. Ainda, com o uso do DTS, é possível selecionar as colunas que podem atravessar o processo de mineração. Esta ferramenta vem junto com o SQL Server 2000 que é a Base de Dados que estamos utilizando, e tem sido bastante utilizada por quem trabalha com Business Intelligence e Data Warehousing.

A construção de Data Marts ou Data Warehouses e a utilização de ferramentas como o DTS não são obrigatórias no processo de KDD, mas podem otimizar e simplificar as etapas de Seleção, Preparação e Limpeza dos dados que podem consumir 80% do tempo gasto com a realização do Processo de Descoberta de Conhecimento.

Em se tratando de ferramentas que apoiam a etapa de Mineração de Dados, algumas são citadas: Intelligent

Miner, Enterprise Miner, MineSet, Clementine, DBMiner [Han & Kamber (2001)].

A ferramenta WEKA (Waikato Enviroment for Knowledge Analysis), [Witten, I.H. & Frank E. (2000)], também tem sido bastante utilizada na realização da etapa de Data Mining. Desenvolvida na linguagem Java, pela Universidade de Waikato na Nova Zelândia, esta ferramenta é de Domínio Público e trabalha com diversas Técnicas de Data Mining.

Ela é composta de dois pacotes que podem ser embutidos em outros programas escritos em Java, permitindo que um desenvolvedor possa criar seu próprio Data Mining Explorer.

Umas das técnicas implementadas no WEKA são as regras de associação, cujo objetivo é encontrar os itens que ocorrem de forma conjunta no Banco de Dados, e formar regras a partir desses conjuntos para que estas sejam utilizadas por um analista para geração de novos conhecimentos. Regras são representadas pela notação X ==> Y (X implica em Y), onde X e Y são conjuntos de itens distintos. Esta implicação é avaliada através dos fatores: suporte e confiança [Viana (2004)] .

Existem vários algoritmos que possibilitam encontrar regras de associação. Entre eles estão: PARTITION, DIC, DHP, DLG, ABS e, o mais comum, o algoritmo APRIORI [IC01 (2004)] .

O algoritmo Apriori na ferramenta Weka trabalha apenas com valores categóricos nominais (strings) e o arquivo que contém os dados deve estar no formato ARFF.

ARFF (Attribute-Relation File Format) é um formato padrão de arquivo texto utilizado para representar datasets. Através do Protótipo implementado, é possível gerar o arquivo ARFF do arquivo Texto exportado através do DTS do SqlServer 2000.

Um arquivo ARFF gerado possui a seguinte estrutura:

- Nome da Tabela

@relation questionario_socio_econômico

- Nome dos campos e seu tipo

@attribute idade {A,B,C,D,E}

@attribute sexo {A,B,C}

- Dados dos campos separados por vírgula

@data

A,A

A,B

3 Testes e Resultados

Neste trabalho, considerou-se como domínio o Data Mart que armazena os dados pertinentes ao 1º Processo Seletivo de 2004 do Centro Universitário em questão.

Este Data Mart (Figura 01) é resultado da integração de algumas tabelas que constituem a base operacional do Processo Seletivo e base de informações acadêmicas sobre alunos matriculados.

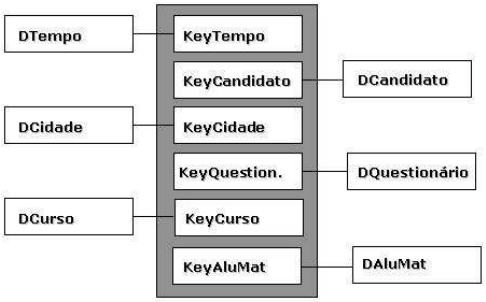

Figura 1: Data Mart construído

A principal tabela a ser utilizada nesta etapa do trabalho é a tabela Dimensão "Questionário". Ela é composta por 42 atributos que são as perguntas descritas no Questionário Sócio-Econômico-Cultural. Este Questionário é entregue ao candidato no período dos processos seletivos através de um gabarito próprio. As respostas dos gabaritos são transformadas em dados digitais através da Leitora Óptica e armazenados em uma Base de Dados SqlServer 2000.

Os dados contidos no Data Mart foram limpos e formatados de acordo com as especificações da Etapa de Pré-Processamento do KDD. Esta etapa é muito importante para que ruídos, ou seja, informações desnecessárias não interfiram no resultado final.

Todos os atributos da Tabela Dimensão "Questionário" que continham menos de 5 opções de respostas tiveram que ser adequadas, como é mostrado nas duas tabelas abaixo:

Tabela 1: Dados Operacionais

A	561
B	415
C	12
D	15
E	11
Z	107

Tabela 2: Dados Transformados

A	561
B	415
Z	145

A opção referente ao Sexo, cujas respostas eram somente A(Feminino) ou B(Masculino), possuía a

Relação Quantitativa de respostas relacionada na Tabela 01.

Os valores referentes às letras C, D, E correspondem às marcações incorretas ou rasuras no gabarito. A letra Z se refere às marcações nulas ou ausentes. Os valores foram então adequados como sendo opções nulas, incorretas ou deixadas em branco, correspondentes à letra Z (Tabela 02).

Algumas tabelas que compõem o Data Mart tiveram alguns atributos eliminados por não contribuírem para o objetivo da análise ou por apresentarem valores omissos (Tabela 03).

Tabela: Dcandidato	
Nome	ELIMINADO – Dado não contribui para o objetivo da análise
Ano	ELIMINADO - Na tabela Dquestionário já existe a questão Idade

Tabela 3 - Exemplo de alguns dos atributos que foram eliminados ou corrigidos

Através do DTS, foram selecionadas as colunas que poderiam ser submetidas ao processo de Mineração dos Dados. Essas colunas foram formatadas e exportadas para o Word onde foram convertidas para o formato do Arquivo ARFF.

O tipo de conhecimento esperado com a realização deste trabalho é a possibilidade de analisar o perfil dos candidatos ao Processo Seletivo, além de encontrar relações interessantes sobre este perfil.

Inicialmente, estando dentro do Data Mart criado, selecionaram-se os atributos necessários para gerar regras que pudessem ser analisadas e, a partir delas, este "perfil" pudesse ser caracterizado.

Em testes iniciais, foram selecionados os atributos referentes às Questões 04, 17, 18 e 39 do Questionário Sócio-Econômico-Cultural:

04) Onde reside atualmente?

17) Qual o fator principal que o levou a escolher a Instituição ?

18) Qual a renda total de sua família ?

39) Como obteve informação sobre o Processo Seletivo do Unifor-MG ?

Utilizando o Assistente DTS (Figura 02), selecionou-se a base no SQL Server que continha os dados a serem exportados. Em seguida, definiu-se o formato do arquivo texto no qual os dados seriam gravados.

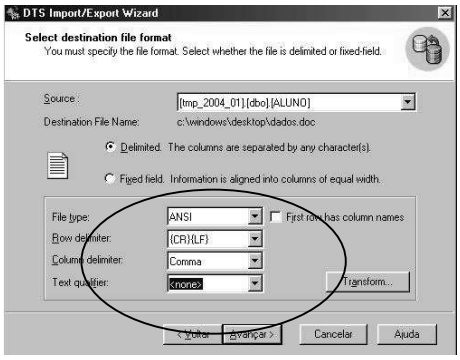

Figura 2: Assistente DTS

Os dados a serem exportados são o resultado da Query anteriomente definida através da seleção de atributos na Tabela Dimensão Dquestionário.

O arquivo texto resultante é transformado em um Arquivo Arff através do protótipo desenvolvido em Java. Este protótipo implementado está sendo utilizado para auxiliar e facilitar a etapa de mineração do trabalho. Ele possui as etapas de Geração do Arquivo Arff (Figura 03), além da etapa de mineração dos dados (Figura 04). Para a etapa de mineração, foram utilizados as classes e os algoritmos disponibilizados na ferramenta WEKA, pois estão construídos na linguagem java, são de domínio público e bastante citados nas literaturas relacionadas ao trabalho. Deste modo, foi possível construir um "Data Mining Explorer" de acordo com as nossas necessidades.

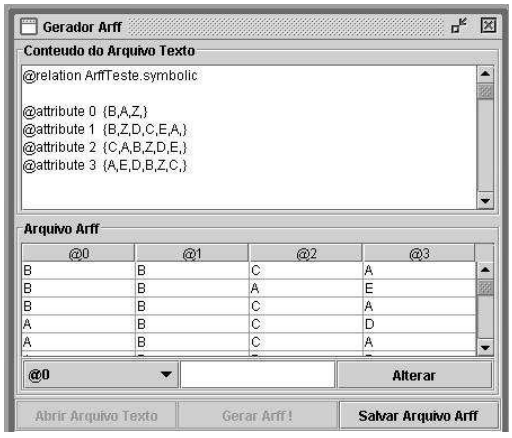

Figura 3 : Gerador Arff

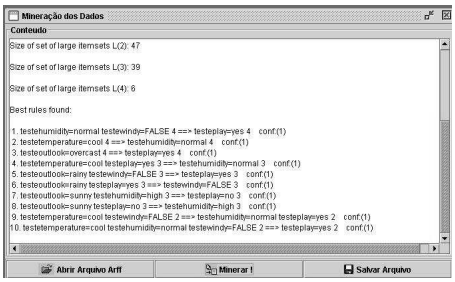

Figura 4 : Etapa de Mineração

O seguinte Arquivo Arff, constituído de 1121 registros e relacionado às 4 questões selecionadas, foi gerado:

```
@relation ArffTeste.symbolic

@attribute onde_residem  {B,Z,D,C,E,A,}
@attribute pq_fuom  {C,A,B,Z,D,E,}
@attribute renda  {A,E,D,B,Z,C,}
@attribute como_soube_vest
{D,A,C,B,Z,E,}

@data
B,C,A,D
B,A,E,A
B,A,D,C
...
```

Apos realizar a etapa de Mineraçao utilizando a técnica de Regras de Associação e o Algoritmo Apriori, as seguintes relações foram geradas:

1.pqfuom= mais_proxima 313 => onde_residem= fga 159 (0.51)
2. pqfuom= mais_proxima 313 => renda=03 a 05 sal 143 (0.46)
3. onde_residem=50Km 267 => renda=3 a 5 m 116 (0.43)
4. onde_residem=100Km 282 =>como_soube_vest=panfleto 121 (0.43)
5. como_soube_vest= panfleto,cartaz 401 =>renda=3 a 5 m 172 (0.43)
6. onde_residem=fga 372 => pqfuom= mais_proxima 159 (0.43)
7. onde_residem=100Km 282 => renda=3 a 5 m118 (0.42)
8. onde_residem=100Km 282 => pqfuom=conceito_da_instituição 116 (0.41)
9. onde_residem=fga 372 ==> renda=3 a 5 m 151 (0.41)

Com essas regras, pode-se observar características interessantes: a quarta regra indica que 43% dos candidatos que moram em uma distância máxima de 100 km de Formiga ficaram sabendo do processo seletivo através de panfletos. Outra relação interessante, indicada pelas regras 1 e 8, mostra que 41% dos candidatos que moram em uma distância máxima de 100 km de Formiga escolheram o UNIFOR devido ao conceito de que desfruta a Instituição, enquanto 51% das pessoas que residem na cidade de Formiga o escolheram por estar mais perto de casa. Através deste resultado, pode-se tentar melhorar a qualidade da divulgação e das informações divulgadas no processo seletivo, atingindo, assim, um público-alvo maior e realmente interessado na qualidade de Ensino do Centro Universitário.

Outro teste realizado foi a seleção dos atributos relacionados ao ensino médio, como o ano de conclusão, o tipo da escola cursada (pública, particular) e o turno cursado, além do sexo e da quantidade de vezes em que o candidato concorreu ao Processo Seletivo da Instituição. As regras geradas foram as seguintes:

1. sexo=Masc QAnoConclEnsMedio=2003 142 => QPrestVestFuom=1ª vez 141 (0.99)
2.sexo=Fem esc_ens_med=estab.publ. QAnoConclEnsMedio=2003 136 => QPrestVestFuom=1ªvez 135 (0.99)
3.esc_ens_med= estab.publ QAnoConclEnsMedio=2003 QTurnoEnsMedio=diurno 128 => QPrestVestFuom=1ªvez 127 (0.99)
4. esc_ens_med= estab.publ QAnoConclEnsMedio=2003 233 => QPrestVestFuom=1ª vez 231 (0.99)
5. QAnoConclEnsMedio=2003 QTurnoEnsMedio=diurno 205 => QPrestVestFuom=1ª vez 203 (0.99)
6.sexo=Fem QAnoConclEnsMedio=2003 190 => QPrestVestFuom=1ª vez 188 (0.99)
7. QAnoConclEnsMedio=2003 342 => QPrestVestFuom=1ª vez 338 (0.99)
8.sexo=Fem QAnoConclEnsMedio=2003 QTurnoEnsMedio=diurno 124 => QPrestVestFuom=1ª vez 122 (0.98)

Analisando todas as relações acima, pode-se concluir que, em um intervalo de 98% à 99%, os candidatos ao processo seletivo realizaram o Ensino Médio em escolas públicas durante o período diurno, o concluíram em 2003 e estavam concorrendo ao Processo Seletivo do Unifor pela primeira vez. Integrando este conhecimento ao conjunto de outras relações geradas, em que a maioria dos candidatos residem com os pais e possuem uma faixa etária entre 17 e 20 anos, em geral, o público atual do processo seletivo de 2004 é bastante jovem.

Pela análise das regras 2, 3, 4, 5, 6 extraídas e exibidas abaixo, pode-se verificar que em um nível de confiança variável entre 54% a 59%, os candidatos que possuem acesso ao computador contam com uma renda superior a 06 salários mínimos. E pela regra 8, verifica-se que os candidatos residentes em Formiga sem acesso ao computador, em 53% das relações exibidas, possuem uma renda total de 03 a 05 salários mínimos. Isto evidencia a diferença da inclusão digital entre os candidatos ao processo seletivo.

1. OndeResidem=Mais de 150 Km 89 ==> RendaTotal=03 a 05 SM 54 conf:(0.61)
2. idade=17 a 18 RendaTotal=01 a 02 SM 63 ==> AcessoAoPC=Não 37 conf:(0.59)
3. idade=19 a 20 Acesso3,AoPC=Não 82 ==> RendaTotal=03 a 05 46 conf:(0.56)
4. RendaTotal=10 a 15 SM 126 ==> AcessoAoPC=Sim, em casa 69 conf:(0.55)
5. RendaTotal=01 a 02 SM 182 ==> AcessoAoPC=Não 99 conf:(0.54)

6. RendaTotal=acima de 15 SM 65 ==> AcessoAoPC=Sim, em casa 35 conf:(0.54)
7. idade=17 a 18 RendaTotal=06 a 09 SM 80 ==> AcessoAoPC=Sim, em casa 43 conf:(0.54)
8. OndeResidem=Formiga AcessoAoPC=não 95 ==> RendaTotal=03 a 05 SM 50 conf:(0.53)

As relações encontradas com a realização do trabalho estão sendo muito válidas e utilizadas pelos responsáveis da Instituição como apoio às tomadas de decisão e na formação de conhecimento útil sobre seu principal cliente, o aluno.

A figura a seguir resume o Processo realizado:

Figura 5 : Processo das etapas realizadas no trabalho

4 Conclusão

A partir dos resultados obtidos com a geração dessas regras de associação, percebe-se que as características dos candidatos em geral são bem diversificadas. Isso pode ser notado analisando o grau de confiança gerado para cada regra que, no geral, é baixo, principalmente no primeiro e terceiro conjuntos de regras geradas e apresentadas na sessão anterior.

Com as experiências adquiridas, pode-se definir que as técnicas de mineração são necessárias e muito úteis, porém, é importante que a visão das possibilidades de utilização e aplicação dessas tecnologias seja ampliada. A partir do conhecimento sobre as regras de negócios da empresa e das técnicas especificadas é possível criar um ambiente de desenvolvimento estratégico para que informações poderosas no apoio às tomadas de decisão possam ser obtidas com êxito.

Como trabalhos futuros, sugere-se a aplicação destas técnicas em outras bases de dados, como a financeira e a acadêmica, visando encontrar relações e conhecimento útil sobre:

- as características e relações do aluno inadimplente;
- o percentual provável de alunos que necessitarão de bolsas de estudos e quais as características destes alunos;
- o histórico escolar e financeiro dos alunos no decorrer dos períodos e prováveis situações e/ou razões de cunho positivo ou negativo que influenciaram neste histórico.

Desta forma, Instituições de Ensino, juntamente com seus cursos na área da Computação, podem contribuir e muito para este novo universo que faz uso da Informação e das Tecnologias de Informação que, cada vez mais, mudarão as relações de competitividade no setor educacional e nos demais segmentos da economia mundial.

5 Referências

[Carvalho(1999)] Carvalho, D. R. *Data Mining através de indução de Regras e Algoritmos Genéticos*. Dissertação de Mestrado em Informática Aplicada, PUCPR, PR, 1999.

[Fayyad(1996)] Fayyad,U.M.,G.Piatetsky–Shapiro, P.Smyth. *Knowledge Discovery and Data Mining: Towards a Unifying Framework*. Proceeding of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, august, 1996.

[Han & Kamber (2001)] Han, J., Kamber, M. *Data Mining: Concepts and Techiniques*.1.ed. New York: Morgan Kaufmann, 2001.

[Viana (2004)] Viana, R. *Mineração de Dados: Introdução e Aplicações*. Artigo SQL Magazine, Ed.10, Ano1.

[IC01 (2004)] *Iniciação Científica – Data Mining*. Disponível em <www.inf.aedb.br/datamining>. Acesso em 23/04/2004.

[Italiano & Esteves (2004)] Italiano,I.C.,Esteves,L.A. *Modelagem de Data Warehouses e Data Marts – Parte 1*. SQL Magazine - Ed.13 - Ano1.

[Witten, I.H. & Frank E. (2000)] Witten, I.H.,Frank, E., *Data Mining: practical machine learning tools and techniques with Java implementations*.2000.

[Zambon & Meirelles (2001)] Zambon, A.C.,Meirelles,J.L. *A Evolução do Processo Decisório e as novas Ferramentas de Apoio à Decisão:Data Warehouse, Olap e Data Mining*. Programa de Mestrado em Engenharia de Produção, São Carlos, 2001.

[Pichiliani (2004)] Pichiliani,M. *DTS: Uma ferramenta para facilitar o processo de ETL*. Artigo Disponível em <www.imasters.com.br>. Acesso em 06/07/2004.