Aplicação de Regras de Associação para Mineração de Dados na *Web*

L. M. R. de Vasconcelos *C. L. de Carvalho*

Technical Report - RT-INF_004-04 - Relatório Técnico
November - 2004 - Novembro

Instituto de Informática
Universidade Federal de Goiás
www.inf.ufg.br

Aplicação de Regras de Associação para Mineração de Dados na *Web*

Lívia Maria Rocha de Vasconcelos * lrocha@inf.ufg.br

Cedric Luiz de Carvalho † cedric@inf.ufg.br

Abstract. *This paper discuss the use of Association Rules to get knowledge from Web. It gives an overview of the Knowledge Discovery in DataBases, specially the data mining step. It is also discussed, superficially, the introduction of semantic in the existing Web and its consequences to the data mining processes.*

Keywords: Data Mining, Association Rules, Apriori Algorithm, Web.

Resumo. *Este texto trata do uso de Regras de Associação para a obtenção de conhecimento na Web. É dada uma visão geral do processo de Descoberta de Conhecimento, especialmente a fase de Mineração de Dados. Também se discute, superficialmente, a introdução de semântica na Web atual e seus reflexos no processo de mineração de dados.*

Palavras-Chave: Mineração de dados, Regras de Associação, Algoritmo *Apriori*,*Web*

1 Introdução

A *Web* é a maior fonte de informação disponível nos dias atuais. Entretanto, grande parte desta informação se encontra escondida no meio da gigantesca massa de dados que é disponibilizada. Neste ambiente, é muito importante que o usuário possa contar com ferramentas que busquem estes dados, de forma inteligente e automática, e os transformem em conhecimento útil.

Neste contexto, surge a necessidade de se descobrir correlações, padrões e tendências entre as informações da *Web*: mineração de dados. A análise sobre essa grande quantidade de dados armazenados pode ser feita usando-se técnicas estatísticas, matemáticas ou de reconhecimento de padrões sobre os dados.

A busca por informações relevantes na *Web* ainda é um problema, principalmente devido à forma em que os dados encontram-se disponibilizados. Como a estruturação do seu conteúdo está voltada preferencialmente para o nível de apresentação, isto é, para a visualização por seres humanos, a extração de conhecimento por máquina é complicada.

*Bolsista de Iniciação Científica - CNPQ / GEApIS/INF/UFG

†Orientador - GEApIS/INF/UFG

A estruturação semântica da *Web* é uma possível solução para este problema. Este é o propósito da Web Semântica. Além da introdução de semântica, ferramentas de Inteligência Artificial, notadamente os agentes inteligentes de software, podem ser utilizadas de forma a tornar mais eficientes os processos de busca por informações na *Web*. No decorrer deste texto, Web Semântica, Mineração de Dados e Mineração de Dados na *Web* serão tratados em detalhe.

O restante deste texto se organiza da seguinte forma: a Seção 2 introduz os conceitos associados à Web Semântica; a Seção 3 dá uma visão geral a respeito de Mineração de Dados; a Seção 4 trata uma das principais tarefas de mineração de dados: associação; a Seção 5 traz explicações sobre conceitos importantes para a implementação de Mineração de Dados no contexto da *Web*; finalmente, a Seção 6 apresenta a conclusão, evidenciando a importância do assunto e soluções encontradas para solucionar os problemas existentes na *Web* atual.

2 Web Semântica

O acesso à informação na *Web*, através de mecanismos automatizados, é ainda bastante precário, devido principalmente à forma com que os dados encontram-se estruturados, já que estes estão orientados, em sua maioria, para o nível de apresentação.

A Web Semântica [9] é uma evolução da *Web* atual. Esta evolução se dá pela atribuição de semântica (significado) aos recursos nela disponíveis. Sua implementação permitirá a construção de mecanismos muito mais eficientes do que os atualmente disponíveis para a recuperação de informação.

Na proposta de desenvolvimento da Web Semântica, é sugerida uma arquitetura de 3 camadas de tecnologias e padrões: Camada de Estrutura, Camada de Esquema (Ontologia) e Camada Lógica.

2.1 Camada de Estrutura

A camada de estrutura é responsável pela representação dos dados e de seus significados. Nesta camada é importante a conceitualização de metadados. Estes podem ser considerados dados que descrevem o conteúdo, a estrutura, a representação e o contexto de um conjunto de dados.

2.1.1 Metadados

Conforme já descrito anteriormente, os metadados podem ser caracterizados como informações estruturadas sobre recursos de informação (artefatos ou serviços) ou mais brevemente como "dados sobre dados". Os metadados são utilizados para colaborar na identificação, descrição, localização e gerenciamento de recursos na *Web*.

Os metadados têm importante papel na representação e troca explícita de informações na *Web*. Eles se dividem em Metadados Estruturais e Metadados Semânticos.

- **Metadados Estruturais**: Os metadados estruturais representam a informação que descrevem a organização e estrutura dos dados gravados, que podem ser, por exemplo, informação do formato, dos tipos de dados gravados e relacionamentos sintáticos entre os dados.

- **Metadados Semânticos**: Os metadados semânticos representam a informação sobre a semântica dos dados, ou seja, sobre seus significados, além do relacionamento entre seus

Tabela 1: Descritores do Padrão Dublin Core [20]

Campos	**Descrição**
Title	Título do recurso
Author	Pessoa ou organização responsável pela criação do conteúdo intelectual do recurso.
Contributor	Pessoa ou organização que contribui intelectualmente na criação do recurso. (Ex.:editor, ilustrador, tradutor, etc).
Publisher	Identifica a entidade responsável por tornar o recurso disponível.
Date	Data da criação ou publicação do recurso.
Source	Informação sobre os recursos que contribuíram para a elaboração do recurso corrente.
Relation	Recursos que possuem relacionamentos com o recurso corrente.
Description	Descrição do conteúdo.
Subject	Tema do recurso.
Type	Forma como o conteúdo é expresso (relatório técnico, dissertação, etc).
Format	Formato em que o recurso é materializado (postScript, HTML, DOC, PDF, etc).
Identifier	Possui o identificador único do recurso.
Language	Idioma.
Coverage	Características espaciais ou temporais.
Rights	Informações sobre os direitos autorais do recurso.

significados. Um exemplo de metadado semântico poderia ser o conteúdo semântico do valor de um dado, como unidades de medida e escala.

Muitos padrões para metadados têm sido definidos nos últimos anos [15], de acordo com áreas específicas para sua aplicação. Um exemplo é o o padrão *Dublin Core* [20], apresentado em uma reunião em Dublin no ano de 1995, com o objetivo de definir uma representação padronizada de metadados para a descrição de recursos eletrônicos. Este padrão é constituído de um conjunto de 15 elementos de meta-informação (Tabela 1).

Para que os dados sejam referenciados corretamente, é necessária sua vinculação a contextos, de forma que seus significados possam ser identificados sem ambigüidades semânticas. As ontologias são um importante componente na diminuição destas ambigüidades.

2.2 Camada de Esquema (Ontologia)

A camada esquema é a responsável por controlar os dados nos documentos, de forma que estes fiquem estruturados e bem definidos quanto ao seu significado. É nesta camada que ocorre a implementação, através do uso de linguagens como XML, RDF e OWL, da definição de relações entre os dados (ontologias).

2.2.1 XML (*eXtensible Markup Language*)

A linguagem XML surgiu como uma alternativa para a falta de estruturação dos documentos, sendo uma tecnologia fundamental para a expansão da Web Semântica. Uma linguagem de marcação, em geral, é um mecanismo para identificar estruturas em um documento. A especificação de XML define um padrão para adicionar marcações no texto [19].

A linguagem HTML, a mais amplamente utilizada na *Web* atual, não é eficiente para o propósito de estruturação dos dados, pois têm um número fixo de *tags*, além de não promover uma separação completa de dados e suas formatações. Esta característica de HTML dificulta a extração dos dados contidos em páginas escritas nesta linguagem, já que os mesmos estão misturados com informações de formatação.

2.2.2 RDF (*Resource Description Framework*)

O RDF é uma recomendação do W3C (*World Wide Web Consortium*) para a definição e o uso de metadados. O RDF é a base para o processamento de metadados e providencia a interoperabilidade entre aplicações, que trocam informações compreensíveis na *Web*. O uso de metadados na *Web*, expressos em RDF, pode facilitar muito a recuperação e o processamento de dados por máquina.

Para melhor compreender RDF, é necessário familiaridade com os conceitos a seguir:

- ***Resource* (Recurso)**: Recurso é tudo que é descrito através de expressões RDF. Um recurso pode ser um documento eletrônico, uma coleção, uma página *Web*, entre outros. Os recursos são sempre nomeados através de URIs (*Uniform Resource Identifier*) que permitem a introdução de identificadores para qualquer entidade.
- ***Property* (Propriedade)**: É um atributo ou característica que descreve o recurso. Uma propriedade representa também o relacionamento entre recursos.
- ***Statement* (Expressão/Afirmação)**: Corresponde à associação de um recurso específico, uma propriedade e um valor desta propriedade para este recurso. Uma declaração é dividida em sujeito, predicado e objeto. O objeto desta declaração pode ser um recurso ou um literal.

2.2.3 Ontologia

Uma ontologia fornece uma conceitualização de um domínio específico de problema, fornece um acordo comum de vocabulários para que os dados sejam referenciados. Assim, serve como uma base comum para a representação de dados e metadados.

Uma ontologia agrega uma larga parte de domínio de conhecimento para cada área em particular e uma das principais motivações para sua construção é a possibilidade de compartilhar e reutilizar conhecimentos, definidos de uma forma genérica, entre comunidades e aplicações. Para que as ontologias possam desempenhar seu papel de integrar semântica à *Web*, faz-se necessário que cada documento disponibilizado na rede possua metadados descritos sob a padronização de alguma ontologia (uma ontologia pode ser representada como uma hierarquia de conceitos[4]).

Os domínios de conhecimento especificados numa ontologia, quando aplicados à recuperação de informação, têm a finalidade básica de servir como esquemas conceituais que darão suporte semântico às consultas. Uma grande vantagem desse contexto é que a ontologia provê uma interpretação semântica unificada para diferentes representações de dados semi-estruturados referentes a um mesmo domínio [14].

2.3 Camada Lógica:

A camada lógica é composta por um conjunto de mecanismos que permitem a realização de inferências. Agentes inteligentes podem atuar nesta camada, utilizando regras de inferência,

para localizar, relacionar e processar informações.

As regras de inferência são mecanismos que possibilitam inferir, a partir de asserções válidas, expressões válidas, isto é, consistentes com todas as interpretações possíveis.

Um exemplo de regra de inferência (*Modus Ponens*) pode ser visto a seguir. Dadas as asserções, consideradas verdadeiras:

- Todo homem é mortal,
- Sócrates é homem

Então, pode-se concluir que: **Sócrates é mortal**.

2.4 Agentes Inteligentes

Agentes inteligentes [16] são programas de computador desenvolvidos para auxiliar o usuário na realização de alguma tarefa ou atividade. Dentre as propriedade que estes podem apresentar, destacam-se autonomia, personalização, reatividade e comunicabilidade.

Hoje, a aplicação dos agentes inteligentes pode ser observada na Educação a Distância (EAD), no comércio eletrônico, na medicina, na área de entretenimento (jogos virtuais), entre outras. A maioria das contribuições do paradigma de desenvolvimento de agentes vem da área de Inteligência Artificial.

A *Web*, por seu carater distribuído e heterogêneo, constitui-se em um ambiente particularmente adequado à aplicação da tecnologia de Agentes Inteligentes, especialmente na recuperação de Informações.

Estes agentes podem ser utilizados especialmente em tarefas de Mineração de Dados, conforme discutido nas próximas seções.

3 Mineração de Dados

A mineração de dados ou *data mining* é considerada o núcleo de processo de descoberta de conhecimento em banco de dados (*Knowledge Discovery in Databases* - KDD). Ela consiste no processo de analisar grandes volumes de dados sob diferentes perspectivas, a fim de descobrir informações úteis que normalmente não estão visíveis ou que dificilmente são encontradas.

Segundo Fayyad [10], pioneiro em KDD, "Mineração de Dados é um passo no processo de KDD que consiste na aplicação de análise de dados e algoritmos de descobrimento que produzem uma enumeração de padrões (ou modelos) particular sobre os dados". A Tabela 2 mostra a origem e evolução da mineração de dados, evidenciando o progresso na questão comercial, as tecnologias disponíveis para a realização das tarefas, os principais fornecedores e as características que marcaram cada fase dessa evolução.

3.1 Descrição do Processo de Descoberta de Conhecimento em Bancos de Dados

Segundo Han e Kamber [13], o processo de descoberta de conhecimento em bancos de dados (KDD) é dividido em sete grandes etapas: limpeza dos dados, integração dos dados, seleção dos dados, transformação dos dados, mineração dos dados, avaliação dos modelos encontrados e apresentação do conhecimento adquirido.

Tabela 2: Quadro de evolução da Mineração de Dados [3]

Etapa Evolucionária	**Questão Comercial**	**Tecnologias Disponíveis**	**Fornecedores de Produtos**	**Características**
Coleção de dados (1960s)	"Qual foi minha receita total nos últimos cinco anos?"	Computadores, fitas e discos	IBM, CDC	Retrospectiva, distribuição de dados estática
Acesso a dados (1980s)	"Quais foram as vendas unitárias de São Paulo em março?"	Bancos de dados relacionais (RDBMS), *Structured Query Language* (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospectiva, distribuição de dados dinâmica a nível de registros
Data Warehousing e Suporte à Decisão (1990s)	"Quais foram as vendas unitárias de São Paulo em março? Avalie também Campinas." .	*On-Line Analytical Processing* (OLAP)	Pilot, Comshare, Arbor, Cognos, Microstrategy .	Retrospectiva, distribuição dinâmica de dados múltiplos níveis
Mineração de Dados	"Qual a previsão para as vendas de Campinas no próximo mês? Por quê?"	Algoritmos avançados, computadores multiprocessados, banco de dados massivos.	Pilot, Lockheed, IBM,SGI, e outras (novas empresas)	Prospectiva, distribuição de informação ativa.

A seqüência de passos do KDD se inicia com a **seleção** de um conjunto de dados ou amostra dos dados com os quais o processo de descoberta será realizado. Estes dados poderão passar por uma etapa de **pré-processamento** onde serão tratados problemas como ruídos e dados incompletos. O passo seguinte consiste na tarefa de **transformação** dos dados, onde poderão ser empregadas operações de projeção e redução. Nesta etapa pode-se reduzir o número de variáveis sob consideração ou encontrar representações dos dados que não variam. Após as etapas descritas, os dados estão prontos para serem processados pela principal tarefa dentro de todo o processo: a **mineração**. São aplicados algoritmos, muitas vezes de forma repetitiva, que procuram por padrões e regras escondidos nos dados. Por fim, as informações descobertas são **interpretadas e avaliadas**, muitas vezes na forma de gráficos ou relatórios, selecionando os **conhecimentos** úteis de todo este processo [11].

Uma seqüência natural neste processo de busca da informação poderia ser:

Dados $\longrightarrow$ Informação $\longrightarrow$ Conhecimento $\longrightarrow$ Decisão.

A Figura 1 demonstra esse processo.

Atualmente, a capacidade do ser humano de coletar dados e armazená-los é muito maior do que a sua capacidade de entendê-los, e através deles, obter conhecimento útil. O processamento das informações tornou-se cada vez mais difícil de ser realizado, devido ao grande volume de dados existente. As últimas décadas vêm mostrando a necessidade de um processo

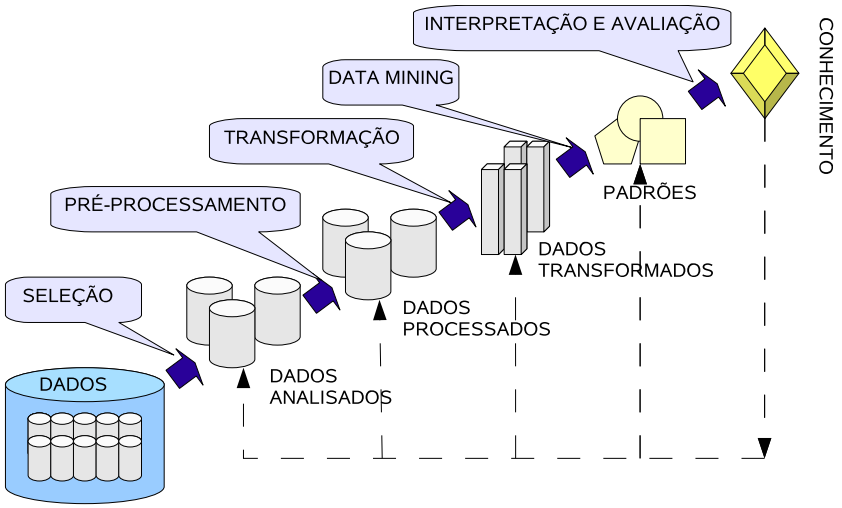

Figura 1: Etapas do processo de descoberta de conhecimento em Bancos de Dados [10].

automatizado para a descoberta de padrões interessantes e desconhecidos em bancos de dados reais. A seguir, serão conceituadas as tarefas mais relevantes no processo de mineração de dados para a descoberta de informações.

3.1.1 Tarefas

Com o notável crescimento da quantidade de dados armazenados em meios magnéticos, a análise individual dos dados torna-se inviável e pouco proveitosa. Neste sentido, surge a necessidade de se extrair informações de conjuntos de dados a partir do uso de tarefas da mineração de dados. As tarefas principais de busca de informação implícita são: Agrupamento, Classificação e Associação.

- **Agrupamento**

  Agrupamento é uma tarefa que procura segmentar populações heterogêneas em subgrupos ou segmentos homogêneos. O processo de formação de grupos de objetos dentro de classes de objetos similares é chamado agrupamento. Um grupo é uma coleção de objetos de dados que são similares a um outro dentro do mesmo grupo [13].

- **Classificação**

  A tarefa de classificação consiste em examinar as características de um objeto e enquadrá-las em conjuntos pré-definidos. Consiste na generalização e especialização de dados que servem para distinguir as classes de modo a predizer dados ou classes de registros não classificados automaticamente [17]. Alguns algoritmos que utilizam os conceitos da tarefa de classificação são: as regras de indução, as árvores de decisão e as redes neurais.

- **Associação**

  A tarefa de associação tem o intuito de identificar associações entre registros de dados que, de alguma maneira, estão ou devem estar relacionados. Sua premissa básica é encontrar elementos que implicam na presença de outros em uma mesma transação [17]. Alguns algoritmos que utilizam os conceitos desta tarefa são as regras de associação e os padrões seqüenciais.

4 Mineração de Dados e Regras de Associação

Diferentes tipos de dados podem ser minerados [5, 7, 13, 21]. Dentre os diversos algoritmos que podem ser utilizados neste processo, como as regras de classificação, padrões seqüenciais e os segmentos de dados (*clusters*), destacam-se, por sua aplicabilidade, as regras de associação. Segundo Brusso [6], "as regras de associação são padrões descritivos que representam a probabilidade de que um conjunto de itens apareça em uma transação visto que outro conjunto está presente".

4.1 Regras de Associação

A tarefa associação tem como premissa básica encontrar elementos que implicam na presença de outros elementos em uma mesma transação, ou seja, encontrar relacionamentos ou padrões freqüentes entre conjuntos de dados. O termo transação indica quais itens foram consultados em uma determinada operação de consulta.

Tipicamente, regras de associação representam padrões existentes em transações armazenadas. Por exemplo, a partir de uma base de dados, na qual registram-se os itens adquiridos por clientes, uma estratégia de mineração, com o uso de regras de associação, poderia gerar a seguinte regra: $\{cinto, bolsa\} \rightarrow \{sapato\}$, a qual indica que o cliente que compra cinto e bolsa, com um determinado grau de certeza, compra também sapato. Este grau de certeza de uma regra é definido por dois índices: o fator de suporte e o fator de confiança.

A tecnologia de "código de barras" permitiu às organizações de varejo coletar e armazenar grande quantidade de dados referentes às compras realizadas por seus clientes, conhecidas como "dados da cesta". Através do conhecimento desses dados, as organizações dirigem seus processos de marketing e promovem estratégias de *layout* e catálogos que possam trazer vantagens a partir dos dados coletados.

As bases de dados envolvidas nestes processos são muito grandes. Assim, é necessário que sejam utilizados algoritmos rápidos e eficientes.

O problema pode ser formalizado como se segue [1]. Seja $I = \{I_1, I_2, I_3, ..., I_m\}$ um conjunto de atributos binários chamados **itens** e seja T uma base de dados de transações, onde cada t é representada por um vetor binário, com $t[k] = 1$ se t indica a compra do item I_k e $t[k] = 0$, caso contrário. Existe uma tupla na base de dados para cada transação. Seja X um conjunto de itens em I. É dito que a transação t **satisfaz** X se, para todos os itens I_k em X, $T[k] = 1$.

Uma regra de associação [2] é uma implicação da forma $X \Longrightarrow Y$, onde $X \subset I$, $Y \subset I$, e $X \cap Y = \oslash$. A regra $X \Longrightarrow Y$ é válida no conjunto de transações T, com o grau de confiança **c**, se **c%** das transações em T que contêm X também contêm Y. A regra $X \Longrightarrow Y$ tem suporte **s** em T, se **s%** das transações em T contêm $X \cup Y$. Se as condições forem satisfeitas, **c%** representará o fator de confiabilidade e **s%** o fator de suporte.

4.2 Decomposição da Tarefa

O problema de se descobrir todas as regras de associação pode ser decomposto em duas partes:

- Encontrar todos os conjuntos de itens que possuam um **suporte** de transações acima de um limite mínimo informado. O suporte para um conjunto de itens é o número das transações que contém este conjunto. São chamados de **conjuntos de itens freqüentes** aqueles que têm suporte igual ou superior ao mínimo estabelecido.

- Gerar as regras de associação a partir dos conjuntos de itens freqüentes. Deve-se selecionar apenas as regras que possuam o grau de confiança mínimo, correspondente à confiança_mínima.

Assim, dado um conjunto de transações, o problema de mineração por regras de associação está em gerar todas as regras que contenham o suporte e confiança iguais ou maiores do que os valores mínimos determinados pelo usuário, referenciados como *suporte_mínimo* e *confiança_mínima*, respectivamente.

O suporte de uma regra $X \Longrightarrow Y$, onde X e Y são conjuntos de itens, é dado pela seguinte fórmula:

$$Suporte = \frac{Freqüência\,de\,X\,e\,Y}{Total\,de\,T} \tag{1}$$

O numerador se refere ao número de transações em que X e Y ocorrem simultaneamente e o denominador ao total de transações.

A sua confiança é dada pela seguinte fórmula:

$$\textit{Confiança} = \frac{Freqüência\ de\ X\ e\ Y}{Freqüência\ de\ X}, \tag{2}$$

O numerador se refere ao número de transações em que X e Y ocorrem simultaneamente. O denominador se refere à quantidade de transações em que o item X ocorre.

O suporte (Equação 1) pode ser descrito como a probabilidade de que uma transação qualquer satisfaça tanto X quanto Y, ao passo que a confiança (Equação 2) é a probabilidade de que uma transação satisfaça Y, dado que ela satisfaz X.

A seguir será dado um exemplo concreto de como as regras de associação funcionam na descoberta de padrões de acesso aos endereços de um servidor *Web*, através da análise de arquivos de *log*. Arquivos de *log* são responsáveis por armazenar informações importantes a respeito dos interesses dos usuários e são obtidos através do registro dessas informações, relativas à buscas realizadas ao longo de uma conexão. A Tabela 3 representa um exemplo de um arquivo de *log* referente a uma busca na base de dados de em um supermercado [11].

As linhas correspondem às transações e as colunas aos itens. E assim, a ausência de um item é representada pelo valor "0" enquanto que sua presença é representada pelo valor "1". Para o exemplo em questão, o *suporte_mínimo* especificado é de 0.3 (30%) e a *confiança_mínima* é 0.8 (80%).

Tabela 3: Entrada de dados para a descoberta de regras de associação.

	Leite	**Café**	**Cerveja**	**Pão**	**Manteiga**	**Arroz**	**Feijão**
Transação 1	0	1	0	1	1	0	0
Transação 2	1	0	1	1	1	0	0
Transação 3	0	1	0	1	1	0	0
Transação 4	1	1	0	1	1	0	0
Transação 5	0	0	1	0	0	0	0
Transação 6	0	0	0	0	1	0	0
Transação 7	0	0	0	1	0	0	0
Transação 8	0	0	0	0	0	0	1
Transação 9	0	0	0	0	0	1	1
Transação 10	0	0	0	0	0	1	0

Considerando-se que X = {CAFÉ} e Y = {PÃO}, pode-se calcular o suporte e a confiança para o conjunto {CAFÉ, PÃO}. Ao ser feita uma análise da Tabela 4, verifica-se que apenas nas transações 1, 3 e 4, estes itens aparecem com valor 1 simultaneamente. Desta forma, o numerador da Equação 1 assume o valor 3, que é o número correspondente à quantidade de transações nas quais estes itens têm valor igual a 1 simultaneamente. O denominador será 10, pois este é o número total de transações ocorridas. Logo, Suporte = 3/10 = 0.3.

O cálculo da confiança para este conjunto de itens é feito da seguinte forma: o numerador da Equação 2 é obtido de maneira semelhante ao numerador da Equação 1, logo seu valor é 3. O denominador equivale à quantidade de transações em que os elementos de X têm valor 1. Assim, o denominador assume o valor 3. Portanto, Confiança = 3 / 3 = 1.

Desta maneira, pode-se obter a regra {CAFÉ} $\Longrightarrow$ {PÃO}, com suporte e confiança iguais ou superiores ao mínimo especificado. O processo para a obtenção de regras como esta é discutino da subseção a seguir.

Um dos algoritmos mais utilizados para a construção do conjunto de itens frenqüentes, o algoritmo Apriori, é discutido a seguir.

4.2.1 Algoritmo Apriori

O algoritmo Apriori [1] (Algoritmo 1) é um dos algoritmos mais conhecidos para mineração por regras de associação. O algoritmo emprega busca em profundidade e gera conjuntos de itens candidatos (padrões) de k elementos a partir de conjuntos de itens de $k-1$ elementos. Os padrões não freqüentes são eliminados. Toda a base de dados é rastreada e os conjuntos de itens freqüentes são obtidos a partir dos conjuntos de itens candidatos.

Algoritmo 1: Algoritmo Apriori

```
   F_1 ←{Conjuntos de itens freqüentes de tamanho 1} /* Na
       primeira passagem k = 1                        */
 1 para k = 2; F_{k-1} != vazio; k++ faça
   |  /* Na segunda passagem k = 2                     */
 2 |  C_k ← apriori-gen (F_{k-1}) /* Novos candidatos  */
 3 |  para todo transação t ∈ T faça
 4 |  |  C_t ← subconjunto (C_k, t) /* Candidatos contidos
   |  |      em t                                     */
 5 |  |  para todo candidato c ∈ C_t faça
 6 |  |  |  c.contagem++
 7 |  |  fim
 8 |  |  F_k ← {c ∈ C_k|c.contagem ≥ MinSup}
 9 |  fim
10 fim
11 Resposta F ← Reunião de todos os F_k
```

1) F_k - conjunto de itens freqüentes de tamanho k (conjunto com k elementos) que atende o suporte mínimo estabelecido. Cada membro deste conjunto tem dois campos. O primeiro é conjunto de itens e o segundo é um contador para o suporte.

2) C_k - Conjunto de itens candidatos de tamanho k. Cada membro deste conjunto tem dois campos. O primeiro é conjunto de itens e o segundo é um contador para o suporte.

A seguir o algoritmo *Apriori* é tratado mais detalhadamente.

4.2.2 Explicação do Algoritmo

O algoritmo principal (Apriori) faz uso de duas subrotinas: *apriori-gen*, para gerar o conjunto de itens candidatos (conjunto composto pelos valores correspondentes ao suporte de cada item). Neste conjunto são considerados todos os itens, independente deles atenderem o suporte_mínimo especificado) e eliminar aqueles que não são freqüentes, e a subrotina *subconjunto*, utilizada para extrair as regras de associação. De forma geral, a sua meta é procurar por relações entre os dados enquanto eles são separados. Simultaneamente, o algoritmo calcula o valor correspondente à confiança e ao suporte.

O algoritmo trabalha sobre uma base de transações em busca de itens freqüentes, ou seja, aqueles que possuem suporte maior ou igual ao suporte mínimo. Desta forma, como entrada, é necessário fornecer um valor correspondente ao suporte_mínimo e outro correspondente à confiança_mínima, além de um arquivo de itens e transações.

O algoritmo é executado da seguinte forma:

- Na primeira passagem, o suporte para cada item individual (conjuntos-de-1-item) é contado e todos aqueles que satisfazem o suporte_mínimo são selecionados, constituindo-se os conjuntos-de-1-item freqüentes (F_1).
- Na segunda iteração, conjuntos-de-2-itens candidatos são gerados pela junção dos conjuntos-de-1-item (a junção é feita através da função *apriori-gen*) e seus suportes são determinados pela pesquisa no banco de dados, sendo, assim, encontrados os conjuntos-de-2-itens freqüentes.
- O algoritmo Apriori prossegue iterativamente, até que o conjunto-de-k-itens encontrado seja um conjunto vazio.

A função **apriori-gen** toma como argumento F_{k-1} (conjuntos-de-(k-1)-itens) e retorna o conjunto dos conjuntos de todos os conjuntos-de-k-itens. Primeiramente, no passo de junção, os elementos de F_{k-1} são combinados, de acordo com o Algoritmo 2.

Algoritmo 2: Função Apriori-gen: passo de junção

inserirEm C_k
selecione $p.item_1, p.item_2, ..., p.item_{k-1}, q.item_{k-1}$
de $F_{k-1}p, F_{k-1}q$
onde $p.item_1 = q.item_1, ..., p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$

A seguir, ocorre o passo de poda, todos os conjuntos de itens $c \in C_k$, tal que algum conjunto-de-(k-1)-itens de c não está em F_{k-1}, são eliminados, de acordo com o Algoritmo 3.

Algoritmo 3: Função Apriori-gen: passo de poda

para todo *conjuntos de itens* $c \in C_k$ **faça**
  para todo *subconjuntos-de-(k-1)-itens* s *de* c **faça**
    se $s \notin F_{k-1}$ **então**
      remova c **de** C_k
    fim
  fim
fim

A seguir, será dado um exemplo prático da execução do algoritmo tendo como base de dados as transações presentes na Tabela 4, onde ocorrem 5 itens. O valor "1" corresponde à presença do item em determinada transação e o valor "0"" corresponde à ausência do mesmo. Para o exemplo em questão, o suporte_mínimo considerado é de 0.2 (20%).

Tabela 4: Transações efetuadas na Base de Dados.

	Item 1	Item 2	Item 3	Item 4	Item 5
Transação 1	1	1	0	0	1
Transação 2	0	1	0	1	0
Transação 3	0	1	1	0	0
Transação 4	1	1	0	1	0
Transação 5	1	0	1	0	0
Transação 6	0	1	1	0	0
Transação 7	1	0	1	0	0
Transação 8	1	1	1	0	1
Transação 9	1	1	1	0	0
Transação 10	0	0	0	0	0

A Tabela 5 é formada pelo conjunto de itens candidatos, em que k = 1, ou seja, cada conjunto possui apenas 1 elemento. A segunda coluna desta tabela contém os valores dos suporte correspondentes a cada item. Como exemplo, para o Conjunto {1}, o suporte é calculado por 6/10, onde 6 é o número de transações em que o Item 1 está presente, e 10 é o total de transações.

Tabela 5: Conjunto de itens de tamanho 1

Conjunto	Suporte
1	0.6
2	0.7
3	0.6
4	0.2
5	0.2

Tabela 6: Conjunto de itens de tamanho 1 que atendem o suporte mínimo

Conjunto	Suporte
1	0.6
2	0.7
3	0.6
4	0.2
5	0.2

A Tabela 6 representa o conjunto F_1, o qual é formado selecionando-se, dentre os conjuntos candidatos em C_1, aqueles que atendem o suporte_mínimo especificado.

O conjunto C_2 (Tabela 7) será formado a partir da junção (combinação) dos itens de F_1. No algoritmo, a combinação entre os itens é feita pela função *apriori_gen*. Esta junção se dá gerando-se combinações dos elementos de F_1, dois a dois, onde o primeiro deles é sempre menor que o segundo. O suporte para os novos conjuntos formados são obtidos pela aplicação da Equação 1 aos dados da Tabela 4. Para o caso do conjunto {1,2}, o valor do suporte é 4 (número de transações que contém os Itens 1 e 2 simultaneamente), dividido por 10 (número total de transações). Portanto, o suporte para este conjunto é 0.4.

Os elementos de C_2 que não obtiverem suporte mínimo não serão considerados freqüentes e, portanto, não farão parte do conjunto F_2 (Tabela 8).

A partir de F_2 será formado o conjunto C_3 (Tabela 9), pela aplicação da função *apriori_gen* que produzirá conjuntos de tamanho 3. Deve ser observado que esta função gera novos

Tabela 7: Conjunto de itens de tamanho 2

Conjunto	Suporte
1, 2	0.4
1, 3	0.4
1, 4	0.1
1, 5	0.2
2, 3	0.4
2, 4	0.2
2, 5	0.2
3, 4	0
3, 5	0.1
4, 5	0

Tabela 8: Conjunto de itens de tamanho 2 que atendem o suporte mínimo

Conjunto	Suporte
1, 2	0.4
1, 3	0.4
1, 5	0.2
2, 3	0.4
2, 4	0.2
2, 5	0.2

conjuntos a partir de dois conjuntos de dois elementos nos quais os primeiros elementos são iguais e o segundo elemento do segundo par é maior que o segundo elemento do primeiro par. Por exemplo, os pares {1,2} e {1,3} geram a tripla {1,2,3} e os pares {1,2} e {1,5} geram a tripla {1,2,5}. Os conjuntos {2,3,4} e {2,3,5} são eliminados no passo de poda, uma vez que os conjuntos {3,4} e {3,5} não fazem parte de F_2.

Tabela 9: Conjunto de itens de tamanho 3

Conjunto	Suporte
1, 2, 3	0.2
1, 2, 5	0.2
2, 3, 4	0
2, 4, 5	0

Tabela 10: Conjunto de itens de tamanho 3 que atendem o suporte mínimo

Conjunto	Suporte
1, 2, 3	0.2
1, 2, 5	0.2

Em seguida gera-se C_4 (Tabela 11), os valores dos suportes são verificados e forma-se o conjuntoF_4 que, no caso, é um conjunto vazio, pois o conjunto C_4 {1,2,3,5} não possui o suporte mínimo exigido.

Tabela 11: Conjunto de itens de tamanho 4

Conjunto	Suporte
1, 2, 3, 5	0.1

4.2.3 Obtenção das Regras de Associação

Após a determinação dos conjuntos de itens freqüentes F, pode-se obter as regras de associação. Deve-se observar que cada item freqüente é um conjunto de k itens. Para cada item freqüente $Y = \{I_1 I_2 ... I_k\}$ de F, com $k \geq 2$, pode-se gerar todas as regras (no máximo k regras) que usam itens do conjunto $I_1, I_2, ..., I_k$. O antecedente de cada uma das regras será o subconjunto X de Y tal que X tem $k-1$ itens, e o conseqüente será o item $Y - X$.

Considerando-se novamente o exemplo apresentado anteriormente, pode-se gerar as regras de associação referentes aos conjuntos obtidos. Para isto, é necessário encontrar quais conjuntos de itens atendem a confiança mínima. Para o exemplo em questão, será considerada uma confiança mínima de 50% ou 0,5.

Realizando-se o cálculo da confiança para os conjuntos de itens freqüentes, tem-se que todos os valores do conjunto F_1 possuem confiança igual a 1 (a fórmula para o cálculo da confiança foi descrita na Subeção 4.2), logo atendem a confiança mínima estabelecida. As regras referentes são triviais, da forma $A \rightarrow A$. Para o conjunto F_2 e para o conjunto F_3, tem-se regras de associação da Tabela 12, que atendem ao grau de confiança mínima.

Tabela 12: Regras de Associação referentes aos conjuntos que atendem a confiança mínima

L	Regras	Confiança
1, 2	$1 \rightarrow 2$	0,67
1, 3	$1 \rightarrow 3$	0,67
2, 3	$2 \rightarrow 3$	0,57

A próxima seção apresentará, detalhadamente, o tema Mineração de Dados na *Web*.

5 Mineração de Dados na *Web*

A *Web* é, atualmente, a maior fonte de informação eletrônica disponível. A abundância de informações e recursos que ela oferece, estimulou o desenvolvimento de ferramentas automáticas para a obtenção de informações, principalmente baseadas em técnicas de mineração de dados.

A mineração de dados na *Web* vem sendo estudada desde meados de 1996, mas tem realmente ganhado importância nestes últimos anos. Ela pode ser conceituada como a descoberta e análise inteligente de informações úteis da *Web*.

Os principais fatores que contribuíram para o crescimento e importância da mineração da Web foram os seguintes:

- O aumento das transações comerciais na *Web*, que motivaram o desenvolvimento de técnicas para a "mineração de uso" (Seção 5.1), pois através delas os sítios de venda puderam identificar perfis dos compradores para montarem melhores estratégias de venda e marketing;
- O desenvolvimento da Web Semântica (Seção 2) e da tecnologia dos agentes da informação, onde as técnicas de mineração na *Web* são utilizadas. Desta forma, os serviços da *Web* poderão tornar-se entidades dotadas de comportamento autônomo e comunicar-se através de uma linguagem comum. A mineração de dados na *Web* será uma ferramenta crucial a ser utilizada pelos agentes e serviços nessa visão da *Web*, pois ela os ajudará em várias tarefas, dentre as quais estão busca por informações, personalização e talvez até como mecanismo de aprendizado.

5.1 Ferramentas de descoberta de conhecimento na *Web*

Para promover a pesquisa e busca de informações na *Web* e o atendimento dos interesses, tanto para os usuários quanto para os criadores e mantenedores da *Web*, vem surgindo, cada vez mais, a necessidade de criação e utilização de ferramentas inteligentes e mais capacitadas para o retorno de conhecimento valioso a quem o procura. Desta forma, estas ferramentas de descoberta de conhecimento na *Web* foram classificadas por Cooley [8], como ferramentas de mineração do conteúdo, de mineração do uso da *Web* ou de mineração da estrutura da *Web*. O conceito específico de cada um destes tipos de ferramenta é fornecido a seguir.

- **Mineração do Uso da *Web***

  A mineração do uso da *Web* pode ser definida como sendo a descoberta automática de padrões de acesso dos usuários aos servidores que disponibilizam informações na rede. Como as organizações constroem os seus sítios da forma que seus projetistas consideram mais apropriada para os seus visitantes, a coleta e posterior análise dos dados referentes aos seus acessos podem esclarecer a natureza do tráfego, auxiliando na compreensão do comportamento dos usuários de forma a verificar se o sítio está eficientemente projetado e organizado.

- **Mineração do Conteúdo da *Web***

  A mineração do conteúdo da *Web* abrange as ferramentas que efetuam recuperação inteligente de informações ou aquelas que abstraem a organização dos dados semi-estruturados contidos na *Web*. Algumas dessas ferramentas fazem uso de agentes inteligentes enquanto outras fazem uso de conceitos baseados em bancos de dados. De uma forma ou de outra é garantida uma busca mais eficiente ou uma estruturação de mais alto-nível dos dados na *Web*.

- **Mineração da Estrutura da *Web***

  Enquanto na mineração do conteúdo da *Web* o interesse se encontra no que há dentro dos documentos, na mineração de estrutura, o interesse está nas informações que existem de forma implícita entre os documentos. Esta categoria de mineração envolve a mineração da estrutura que há por trás da interligação entre os documentos da *Web*. Seu principal foco está nos os vínculos de hipertexto que liga os documentos.

5.1.1 Regras de Associação na *Web*

As regras de associação são o mais novo e mais eficiente entre os tipos comuns de padrões em mineração de dados, possuindo, portanto, muitas aplicações potenciais a serem exploradas. Uma destas aplicações está na análise do registro dos acessos aos servidores que disponibilizam documentos na *Web*.

À medida que os usuários interagem com os portais são fornecidos dados sobre estes usuários e sobre como eles respondem ao conteúdo oferecido [12]. A partir destes dados, pode-se descobrir de onde eles vêm, quais páginas visitaram, quando e quanto tempo dispenderam na visita, etc. Pode-se ainda promover a recuperação de informações, através da mineração de conteúdo da *Web*. Assim, estes dados podem ser coletados, seja através do arquivo de *log* convencional do servidor HTTP ou por meio de mecanismos alternativos. Com o passar do tempo, é gerado um volume considerável de dados que podem auxiliar na compreensão do comportamento dos usuários e na melhor organização e estruturação dos recursos oferecidos aos mesmos.

Existe disponível uma grande quantidade de ferramentas, tanto comercialmente como de domínio público, para a análise estatística do acesso às páginas hospedadas em um servidor [18]. Estas ferramentas oferecem informações como contagem de acessos por página, por dia da semana ou do mês, volume trafegado, entre outros. Devido às características dos hiperdocumentos que estão disponibilizados na *Web*, onde cada usuário pode optar por uma série de alternativas para a navegação e interagir de forma pouco previsível, estas estatísticas simples não possuem a profundidade necessária para completa percepção da utilização do servidor [22] e a compreensão do perfil dos usuários.

5.1.2 Exemplo do uso de Regras de Associação

Sistemas de comércio eletrônico, normalmente, usam páginas da *Web* como vitrines para seus produtos e serviços. Clientes podem visitá-las e realizar transações, como comprar produtos e/ou servicos. A mineração dos dados coletados por estes sistemas de comércio eletrônico pode fornecer grande quantidade de informações valiosas originadas do comportamento do consumidor e a qualidade das estratégias de negócio pode ser melhorada.

Por exemplo, a partir da mineração de um repositório de dados, pode-se extrair regras de associação como as seguintes:

- 55% dos usuários que acessaram a página `www.inf.ufg.br` também acessaram a página `www.ufg.br`;
- 30% dos usuários que acessaram a página `www.lojax.com.br/promoções.html` compraram um produto na página `www.lojax.com.br/pedido.html`.

Um outro exemplo interessante é o uso de regras de associação em sistemas para ensino a distância. Pode-se, por exemplo, identificar padrões comportamentais de alunos em um curso de ensino a distância tais como: 70% dos alunos que permaneceram um tempo acima de Ts em uma mesma página X, em seguida, consultaram a página Y; ou alunos que consultaram a página do dicionário de inglês mais que 30 vezes no dia acabaram por desistir do curso.

5.1.3 Mineração de Dados e a Web Semântica

Os resultados dos processos de mineração de dados podem ser muito ampliados quando esta é aplicada sobre dados qualificados por metadados. Neste caso, o uso de contagens (estatísticas), elemento chave na mineração de dados pode ser muito enriquecido com o acréscimo de filtros semânticos.

O uso de filtros semânticos permite a obtenção de regras de associação mais direcionadas para contextos específicos.

Como se pode ver na Figura 1, a mineração de dados é apenas um passo na Descoberta do Conhecimento. A Web Semântica pode ser considerada com um grande "armazém de dados" (*datawarehouse*). A semântica aplicada aos dados, por meio de metadados e ontologias, permite que as fases de **seleção**, **pré-processamento** e **transformação** possa ser automaticamente tratada por mecanismos, como agentes inteligentes, de forma a tornar todo o processo muito mais simples e eficiente.

6 Conclusão

A partir do que foi exposto, pode-se perceber que a Web atual possui uma grande quantidade de problemas, pelo fato de guardar um enorme volume de dados que não se encontram bem estruturados. Por este motivo, a análise automatizada das informações que estão contidas nestes dados é bastante complicada.

Como a capacidade de armazenar dados excede a capacidade de recuperar informações, surge a necessidade de estruturar os dados de forma que eles não se voltem apenas para o nível de apresentação, ou seja, é importante atribuir significados a eles. Esta é a proposta da Web Semântica, cuja intenção principal é associar semântica aos dados contidos na *Web*.

Para que a *Web* atual se estenda com a implantação da Web Semântica, é necessário promover a estruturação dos seus dados através do uso de linguagens como XML, RDF e OWL,

que são consideradas as mais eficientes neste contexto. Vale salientar que metadados e ontologias são imprescindíveis nesta construção, por se tratarem de conceitos básicos e de grande relevância para a Web Semântica.

Com a efetiva implantação de uma *Web* onde os recursos estão semanticamente especificados, a descoberta de conhecimento neste ambiente torna-se-á muito mais facilitada, uma vez que mecanismos automatizados poderão ser utilizados de forma mais eficiente. Desta forma, a Mineração de Dados, um processo interativo que utiliza técnicas de análise de dados para descobrir padrões e relacionamentos entre esses dados poderá produzir resultados muito mais precisos.

As Regras de Associação, com a aplicação do algoritmo *Apriori* são ferramentas muito úteis, que têm sido amplamente aplicadas no contexto de dados estruturados, podem também ser aplicadas sobre dados armazenados em formatos não homogêneos (estruturados, semi-estruturados, ou não estruturados) como ocorre na *Web*, de forma a produzir relações entre os diversos recursos disponíveis nesse ambiente. Com a especificação semântica destes recursos os processos de preparação dos dados para a mineração tornam-se simplificados, uma vez que esta especificação, de certa forma, uniformiza a descrição destes recursos.

7 Agradecimento

À Profa. Dra. Ana Paula Laboissière Ambrósio pela avaliação do presente texto e pelas sugestões feitas, as quais muito contribuiram para a melhoria do texto original.

Referências

[1] AGRAWAL, R; IMIELINSKI, T; SWAMI, A. **Mining Association Rules between Sets of Items in Large Databases.** In: ACM SIGMOD CONFERENCE ON MANEGEMENT OF DATA, p. 207 – 216, Washington, DC, USA, 1993. ACM Press - New York, NY, USA.

[2] AGRAWAL, R; SRIKANT, R. **Fast Algorithms for Mining Association Rules.** In: Bocca, J. B; Jarke, M; Zaniolo, C, editors, PROC. 20TH INT. CONF. VERY LARGE DATA BASES, VLDB, p. 487–499, Washington, DC, USA, 12–15 1994. Morgan Kaufmann.

[3] AYROSA, P. P. D. S. **Fundamentos de Inteligência Artificial.** `http://www.dc.uel.br/~vhmanfredini/ia/DMining/Data\_mining.htm`, acessado em Janeiro de 2003, 2003.

[4] BENJAMINS, V; FENSEL, D; DECKER, S; PEREZ, A. G. **(KA)2: Builiding Ontologies for the Internet.** Journal of Human-Computer Studies (IJHCS), 51(1):687–712, 1999.

[5] BERRY, M. J; LINOFF, G. **Data Mining Techniques: For Marketing, Sales, and Customer Support**. John Wiley & Sons, Inc., New York, NY, USA, 1997.

[6] BRUSSO, M. J. **Access Miner: Uma proposta para a Extração de Regras de Associação Aplicada à Mineração do Uso da *Web*.** Master's thesis, PPGC da UFRGS, Porto Alegre - RS, 2000.

[7] CABENA, P; HADJINIAN, P.AND STADLER, R; VERHEES, J; ZANASI, A. **Discovering Data Mining: From Concept to Implementation.** Prentice Hall, 1997.

[8] COOLEY, R; MOBASHER, B; SRIVASTAVA, J. **Web Mining: information and pattern Discovery on the World Wide Web.** In: 9TH IEEE INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE (ICTAI 97)., p. 558–567, Newport Beach, CA, USA, 1997.

[9] DE LIMA, J. C; DE CARVALHO, C. L. **Uma Visão da Web Semântica**. Technical Report INF_001/94, Instituto de Informática - Universidade Federal de Goiás, Março 2004. Disponível em `http://www.inf.ufg.br/virtualbib/RT-INF_001-04.pdf`.

[10] FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **The KDD Process for Extracting Useful Knowledge from Volumes of Data.** Communications of the ACM, 39(11):27–34, November 1996.

[11] FREITAS, A. A; LAVINGTON, S. H. **Mining Very Large Databases with Parallel Processing.** The Kluwer international series on advances in database systems. Kluwer Academic Publishers, Boston, 1998.

[12] GREENING, D. R. **Data Mining on the *Web*.** Web Techniques, 5:41–46, Janeiro 2000.

[13] HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques.** The Morgan Kaufmann series in data management systems. Morgan Kaufmann Publishers, San Francisco, USA, August 2001.

[14] MELLO, R. D. S; DORNELES, C. F; KADE, A; BRAGANHOLO, V. D. P; HEUSER, C. A. **Dados Semi-Estruturados.** In: XV SIMPóSIO BRASILEIRO DE BANCO DE DADOS / XIV SIMPóSIO BRASILEIRO DE ENGENHARIA DE SOFTWARE., p. 475–513, 2000.

[15] MORI, A; DE CARVALHO, C. L. **Metadados no Contexto da Web Semântica**. Technical Report INF_002/94, Instituto de Informática - Universidade Federal de Goiás, Março 2004. Disponível em `http://www.inf.ufg.br/virtualbib/RT-INF_002-04.pdf`.

[16] RUSSELL, P; STUART, N. **Artificial Intelligence. A Modern Approach**. Prentice-Hall, 15th edition, 2004.

[17] SCHUNEIDER, L. F. **Mineração de Dados - Conceitos.** Universidade Federal do Rio Grande do Sul, UFRS, 2002.

[18] UNIVERSITY, U. **Access Log Analysers.** `http://www.uu.se/Software/Analyzers/Access-analyzers.html`, acessado em Outubro de 2003, 2003.

[19] WALSH, N. **What is XML?** `http://www.xml.com`, acessado em Novembro de 2003, 1998.

[20] WEIBEL, S.AND GODBY, J; MILLER, E; DANIEL, R. **Dublin Core Metadata Initiative.** `http://dublincore.org/`, acessado em Janeiro de 2004, 1995.

[21] WITTEN, I. H; FRANK, E. **Discovering Data Mining: From Concept to Implementation.** Morgan Kaufmann Publishers, San Francisco, Calif., USA, 2000.

[22] ZAIANE, O. R; XIN, M; HAN, J. **Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs.** In: ADVANCES IN DIGITAL LIBRARIES CONFERENCE (ADL 98), p. 19, Washington, DC, USA, 1998. IEEE Computer Society.