Um Modelo para a Descoberta de Regras de Associação Aplicado à Mineração do Uso da *Web*

Marcos J. Brusso[1,2], Philippe O. A. Navaux[1], Cláudio F. R. Geyer[1]

[1] Instituto de Informática da UFRGS
[2] Instituto de Ciências Exatas e Geociências da UPF
{brusso, navaux, geyer}@inf.ufrgs.br

Resumo

As regras de associação (RA's) são padrões descritivos que representam a probabilidade de um conjunto de itens aparecer em uma transação visto que outro conjunto está presente. Dentre as possibilidades de aplicação da mineração de dados na Web, a mineração do seu uso consiste na extração de regras e padrões que descrevam o perfil dos visitantes aos sites e o seu comportamento navegacional. Este trabalho descreve um modelo para a extração de regras de associação aplicado ao uso da Web o qual caracteriza-se por enfocar as etapas do processo de descoberta do conhecimento desde a obtenção dos dados até a apresentação das regras obtidas ao analista, considerando características específicas do domínio.

1 Introdução

Um dos tipos comuns de padrões que podem ser extraídos através da mineração de dados são as regras de associação. Elas representam a probabilidade de que um item apareça em um conjunto, ou transação, dado que outro está presente. Estas regras tem sido utilizadas principalmente no comércio varejista para a análise dos itens adquiridos pelos consumidores em uma cesta de compra. Segundo John [JOH 97], as regras de associação são o mais novo 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 *World Wide Web* (servidores HTTP).

A medida em que os usuários interagem com os sites, são fornecidos dados sobre eles e sobre como eles respondem ao conteúdo oferecido [GRE 00]. Como exemplos destes dados, pode-se descobrir de onde eles vêm, quais páginas visitaram, quando e quanto tempo despenderam na visita. Estes dados podem ser coletados, seja através do arquivo de *log* convencional do servidor HTTP ou por meio de mecanismos alternativos, gerando, com o passar do tempo, 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.

Existem disponíveis uma grande quantidade de ferramentas, tanto comercialmente como de domínio público, para a análise estatística do acesso às páginas hospedadas em um servidor [UPP 99]. Estas ferramentas oferecem informações como contagem de acessos por página, por dia da semana ou do mês, volume trafegado, etc. 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 simples estatísticas não possuem a profundidade necessária para completa percepção da utilização do servidor [ZAI 98] e a compreensão do perfil dos usuários. Neste contexto, surgiu uma nova família de ferramentas que, através da aplicação de algoritmos mais inteligentes como os de mineração de dados, são capazes de extrair conhecimento útil a partir dos acessos dos usuários ao conjunto de páginas de um site. Estas ferramentas foram classificadas por Cooley [COO 97] como sendo de mineração do uso da *Web*.

Alguns trabalhos já foram publicados propondo alternativas para a aplicação de técnicas de mineração no ambiente da *Web* e algumas ferramentas para tal foram construídas [COO 97, SPI 99, ZAI 98] . Contudo, como esta é uma área de pesquisa recente, ainda existem muitos pontos a serem pesquisados de forma a atingir problemas deixados em aberto pelos outros trabalhos ou abordar novas questões pertinentes à este tipo particular de aplicação.

2 O Modelo Proposto

O modelo proposto para o processo de mineração pode ser dividido, de um ponto de vista mais amplo, em cinco etapas distintas, conforme pode ser visualizado na Figura 2.1: a obtenção dos dados, o pré-mineração, a mineração, o pós-mineração e a interpretação dos resultados.

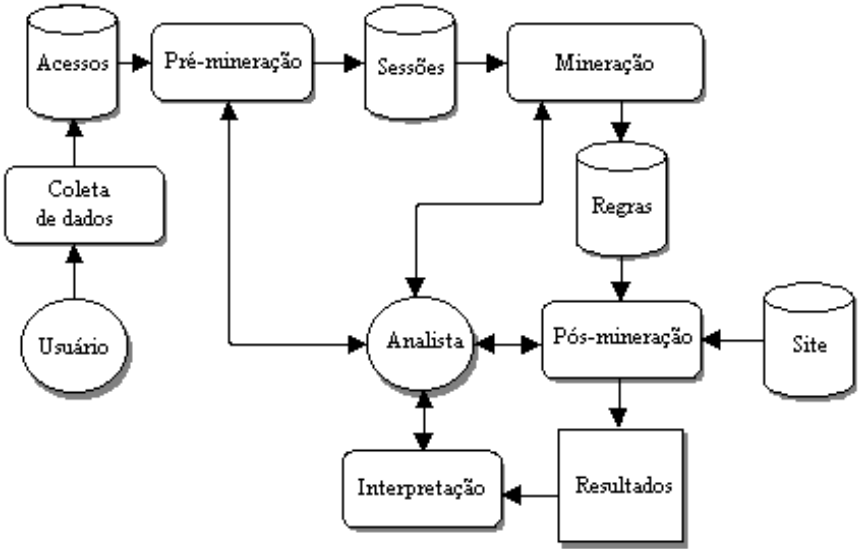

FIGURA 2.1 - Estrutura geral do modelo

A primeira etapa corresponde à coleta e armazenamento dos dados a partir dos acessos dos usuários às páginas de um *site* a fim de formar a base de dados que servirá como origem para todo o processo de mineração. Dessa forma, esta etapa tem como entrada as visitas dos usuários e, como saída, os registros de acessos das mesmas. Devido aos dados incompletos no arquivo de *log* convencional, causado pelo uso de diversos níveis de cache e a dificuldade de identificação das sessões de navegação, o modelo proposto faz uso de um mecanismo alternativo para a obtenção dos registros de acesso, onde cada página que se deseja monitorar é modificada para que seja inserido um *script*, o qual será responsável pela parte do processo a ser executada no cliente. Quando o cliente solicita um documento ao servidor, esse é recebido juntamente com o *script* embutido. No momento em que o documento é carregado no navegador do usuário, o *script* é executado e fica a seu cargo solicitar ao servidor a execução de um processo responsável por armazenar os dados referentes ao acesso, passando como parâmetro a identificação da página que acaba de ser carregada.

A segunda etapa do processo, a pré-mineração, tem como finalidade efetuar todo o tratamento necessário aos dados a fim de torná-los adequados para a etapa de mineração. Essa etapa tem como entrada os registros de acesso armazenados no arquivo de *log* proposto e os critérios especificados pelo analista e, como saída, as sessões dos usuários que atendem aos critérios de seleção definidos, em um formato apropriado para a aplicação do algoritmo de mineração de regras de associação.

A etapa de mineração corresponde à aplicação do algoritmo para a extração de regras de

associação sobre o arquivo que contém as sessões de navegação resultantes da etapa anterior a fim de se obter os referidos padrões. Como entrada para esta etapa, além do conjunto de sessões, também se previu uma série de parâmetros a serem definidos pelo analista; como saída, esta etapa resulta em um conjunto, eventualmente vazio, de regras que coincidem com os critérios especificados. O algoritmo escolhido para a mineração das regras de associação foi o *Apriori* [AGR 96], opção que foi motivada pelo fato de ele ser citado na literatura [GUI 98, OGU 98] como o estado da arte e que seria eficiente, tendo um desempenho superior ao das demais propostas.

A pós-mineração é a etapa responsável pelo tratamento das regras extraídas na etapa anterior antes que elas sejam apresentadas ao analista, a fim de que o trabalho de sua interpretação seja facilitado e mais produtivo. Esse processamento leva em consideração critérios definidos pelo analista, como a seleção de apenas regras que possuam um determinado conjunto de páginas em seu antecedente ou conseqüente, e informações extraídas do domínio, no caso, a estrutura do *site*.

A necessidade de seleção das regras com base na estrutura do site surgiu a partir de análise inicial de alguns resultados obtidos com um protótipo da ferramenta de mineração, onde percebeu-se que muitas regras encontradas, apesar de possuírem um grau de confiança alto, não representavam novo conhecimento [BRU 99a]. Isso se deve ao fato de tais regras apenas descreverem o caminho natural do usuário dentro do conjunto de páginas, o qual é forçado a tal pela própria estrutura de *links* disponibilizada.

Para determinar se uma regra é, provavelmente, sem valor de interesse, deve-se testar a ocorrência do conjunto de páginas que formam o seu conseqüente entre todos caminhos que unem cada subconjunto de duas páginas do seu antecedente. Com exceção da situação em que o antecedente da regra é composto por uma única página, implicando a inexistência de par a analisar, podem ser encontradas quatro situações distintas:

- não existir caminho que una as duas páginas (*a*);
- existir algum caminho que as una, sendo que:
    - nenhum deles possui o conseqüente (*b*);
    - alguns deles possuem o conseqüente (*c*) ;
    - todos eles possuem o conseqüente (*d*).

Uma regra de associação pode ser considerada potencialmente sem interesse se existir, pelo menos, um subconjunto de duas páginas no antecedente da regra, tal que o conseqüente dela está presente em todos os caminhos possíveis entre ambas. Isso equivale a existir algum caso em que a situação (*d*) se aplica.

A etapa final de interpretação dos resultados, que corresponde à análise das regras obtidas pelo especialista no domínio e à decisão de aproveitá-las em seu favor, ou de retornar para alguma das etapas anteriores, por ser baseada em muitos aspectos subjetivos do ponto de vista do analista, não poderia ser automatizada por nenhuma ferramenta, uma vez que não se pode substituir a figura do humano no processo. Dessa forma, esta etapa está fora do escopo deste trabalho.

3 Conclusões

Este trabalho descreveu um modelo para o proceso de extração de regras de associação aplicado a mineração do uso da *Web*. Este ambiente se mostrou propício para a tarefa de mineração de dados, porém o *log* tradicional não é adequado para a descoberta de RA's, de maneira que um mecanismo alternativo para a coleta de dados foi proposto.Uma vez que as medidas

independentes do domínio não são suficientes, a estrutura do *site* sendo analizado foi considerada a fim de selecionar as regras potencialmente mais interessantes.

Referências Bibliográficas

[AGR 96] AGRAWAL, Rakesh et al. Fast Algorithms for Mining Association Rules. In: FAYYAD, Usama M. et al. **Advances in Knowledge Discovery and Data Mining.** Menlo Park: AAAI Press, 1996. 611p. p.307-328.

[BRU 99a] BRUSSO, Marcos José. O Uso de Mineração de Dados na Descoberta do Comportamento do Usuário da Web. In: SEMANA ACADÊMICA DO PPGC. 4.:1999, ago.16-20. Porto Alegre. **Anais...** Porto Alegre: PPGC da UFRGS, 1999. 391p. p.183-186.

[COO 97] COOLEY, Robert; MOBASHER, Bamshad; SRIVASTAVA, Jaideep. Web Mining: Information and Pattern Discovery on the World Wide Web. In: 9th IEEE INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE, 1997, Newport Beach. **Proceedings...** 1997.

[GRE 00] GREENING, Dan R. Data Mining on the Web. **Web Techniques**. San Francisco, v.5, p.41-46. Jan. 2000.

[GUI 98] GUILLAUME, Sylvie; GUILLET, Fabrice; PHILIPPÉ, Jacques. Improving the Discovery of Association Rules with Intensity of Implication. In: PRINCIPLE OF DATA MINING AND KNOWLEDGE DISCOVERY, SECOND EUROPEAN SYMPOSIUM, Nantes, 1998. **Proceedings...** Springer, 1998. p.318-327.

[JOH 97] JOHN, George H. **Enhancements to the Data Mining Process.** Stanford: Stanford University, Ph.D. Dissertation. 1997.

[OGU 98] OGUCHI, Masato et al. Characteristics of a Parallel Data Mining Application Implemented on an ATM Connected PC Cluster. In INTERNATIONAL CONFERENCE AND EXHIBITION ON HIGH-PERFORMANCE COMPUTING AND NETWORKING, 1997. **Proceedings...** p.303-317.

[SPI 99] SPILIOPOULOU, Myra; FAULSTICH, Lukas C. WINKLER, Karsten. A Data Miner analyzing the Navigational Behaviour of Web Users. In: Workshop on Machine Learning in User Modelling of the ACAI'99 Int. Conf., Creta, **Proceedings...** 1999.

[UPP 99] UPPSALA UNIVERSITY. **Access Log Analysers**. Disponível por WWW em http://www.uu.se/Software/Analyzers/Access-analysers.html (12 Mar. 1999).

[ZAI 98] ZAÏANE, Osmar R.; XIN, Man; HAN, Jiawie. Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", Proc. Advances in Digital Libraries Conf. (ADL'98), **Proceedings...** Santa Barbara, 1998. p. 19-29.