Pós-mineração Paralela de Regras de Associação*

Roberto Niche**, Roni Antônio Dall Orsoletta**, Marcos José Brusso, Marcelo Trindade Rebonatto

Universidade de Passo Fundo
BR 285 Caixa Postal 611, Passo Fundo - RS, CEP 99001-970
Fone:(54) 316.8354
{43941, 43945}@lci.upf.tche.br, {brusso, rebonatto}@.upf.br

Introdução

O processo de Descoberta de Conhecimento em Bases de Dados (DCBD) consiste, de acordo com [FAY 96], da utilização de ferramentas computacionais a fim de descobrir informações valiosas, potencialmente úteis, descritas na forma de padrões, a partir dos volumes de dados que estão sendo coletados e armazenados pelas organizações atualmente. A mineração é a etapa responsável pela extração dos padrões, dentro do processo de DCBD, o qual consiste em uma série de etapas iniciando com o pré-processamento dos dados e tendo o pós-processamento dos padrões descobertos como etapa posterior, que por isto, também é chamada de pós-mineração [FEL 98].

A descoberta de regras de associação é uma tarefa de mineração de dados que tem por objetivo encontrar relacionamentos ou padrões freqüentes entre conjuntos de dados. Uma regra de associação é um padrão descritivo que representa uma declaração na forma $X \rightarrow Y$. O conjunto de itens X, que aparece à esquerda do operador de implicação, é denominado antecedente da regra; por sua vez, o conjunto de itens Y, que aparece à direita do operador, é denominado de conseqüente. As técnicas de mineração de regras de associação podem ser aplicadas no contexto da Web para encontrar páginas ou conjuntos de páginas que normalmente são acessadas pelo mesmo usuário em um *site*, sendo que uma transação consiste em uma de suas visitas ao servidor. Como exemplo dessas regras pode-se citar "*80% dos usuários que acessaram a página X.html também acessaram a página Y.html*"

O algoritmo seqüencial de pós-mineração

A pós-mineração é a etapa responsável pelo tratamento das regras extraídas pelas ferramentas de mineração, antes que elas sejam apresentadas ao analista, a fim de que o trabalho de interpretação seja facilitado e mais produtivo. No modelo original, denominado *AccessMiner* [BRU 01], esse processamento leva em consideração informações extraídas da estrutura do *site*. Esta estrutura possui informações incorporadas que podem auxiliar na seleção das regras que são potencialmente mais interessantes, assim como na eliminação daquelas que provavelmente não serão aproveitadas. O estudo realizado por [BRU 00] demonstrou que muitas regras, apesar de

* Apoio FAPERGS e Universidade de Passo Fundo
** Bolsistas PIBIC/UPF

possuírem um grau de confiança alto, não representavam novo conhecimento. Isso se deve ao fato de que tais regras apenas descrevem o caminho natural do usuário dentro do conjunto de páginas, o qual é forçado a tal pela própria estrutura de *links* disponibilizada.

Tomando-se como exemplo um *site* hipotético, ilustrado na Fig. 1, percebe-se que a regra $\{A, E\} \rightarrow B$, ou seja, "*os usuários que visitam as páginas A e E também visitam a página B*", embora possa ter um grau de confiança elevado, simplesmente descreve o que já era de se esperar em razão da estrutura do referido *site*. Isso se deve ao fato de o usuário, para que possa visitar ambas as páginas que compõem o antecedente da regra, ser forçado a fazê-lo, normalmente, utilizando os *links* que passam pela página *B*. Assim, pode-se assumir previamente que essa regra não deve ser interessante. Seguindo essa lógica, deduziu-se que, se o conseqüente da regra aparecer como caminho obrigatório entre as páginas do seu antecedente, então a regra é, provavelmente, sem valor de interesse, o que [BRU 01] chamou de *regra estruturalmente trivial*.

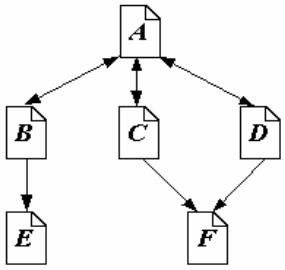

Fig. 1. Estrutura do *site* hipotético

Uma regra de associação é, então, considerada *estruturalmente trivial* 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.

O algoritmo paralelo implementado

A utilização do algoritmo seqüencial para o pós-processamento das regras de associação, separando-as em potencialmente mais interessantes ou triviais, mostra-se uma alternativa adequada para auxiliar na tarefa de análise das regras obtidas na etapa de mineração, segundo [BRU 01]. Porém, o algoritmo é muito oneroso na utilização de recursos computacionais, necessitando de elevado tempo para sua execução.

Após a análise do fluxo de processamento da versão seqüencial do algoritmo de pós-mineração, optou-se pela escolha de um algoritmo paralelo do tipo mestre-escravo. Este tipo de algoritmo adapta-se bem com problemas onde a definição da carga de trabalho a ser resolvida não pode ser determinada estaticamente, como é o caso da pós-mineração de regras de associação. Isto ocorre em virtude das regras de associação possuírem um número não fixo de páginas nos seus antecedentes. Além disto, o número de caminhos possíveis a serem investigados depende da quantidade de *links* disponíveis entre as páginas que compõem cada regra. O processo *mestre* irá controlar a carga de trabalho de cada processador, enviando tarefas e recebendo resultados. Por sua vez, os processos escravos irão efetivamente processar as regras de associação, através da

execução do algoritmo de classificação e enviando os resultados ao processo mestre. O modelo proposto está ilustrado, na Fig. 2.

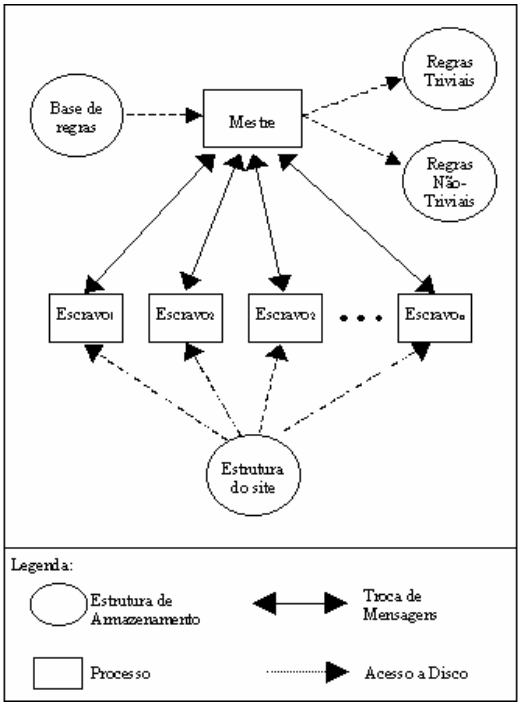

Fig. 2. Modelo Paralelo Proposto

Na divisão das tarefas, o processo mestre irá constantemente acessar a base de regras a serem processadas, enviando-as alternadamente para cada um dos processos escravos, que irão realizar a classificação da regra e devolver o resultado ao processo mestre. Para a tarefa de classificação da regra em trivial ou não, os escravos irão utilizar a estrutura do *site*. Desta forma, esta estrutura é carregada em memória no início da execução, na forma de um grafo, por cada um dos processos escravos. A finalização do processamento acontece no momento em que o processo mestre atinge o final da massa de dados a ser analisada, enviando, nesta ocasião, sinal a todos os processos cooperantes. Têm-se, então, os resultados completos armazenados na máquina responsável pelo processo mestre.

Implementação e resultados

O processo de pós-mineração de regras de associação, tanto a versão seqüencial quanto o algoritmo paralelo apresentado foram implementados utilizando a linguagem de programação C, padrão ANSI. O modelo paralelo foi implementado utilizando LAM/MPI 6.5.6. Na obtenção dos resultados foi utilizado o agregado de computadores do Grupo de Pesquisa em Computação Paralela e Distribuída (ComPaDi/UPF), composto de cinco máquinas, adquirido com auxílio financeiro da FAPERGS.

A base de dados de entrada utilizada para a avaliação dos algoritmos consiste em 55.240 regras de associação obtidas com uso da ferramenta *AccessMiner* [BRU 01], a partir de um total de 257.446 acessos. A estrutura do *site* analisado contém 54 páginas,

cada uma delas tendo, em média, 3,2 *links* para outras páginas do mesmo *site*. A Fig. 3(a) apresenta os tempos de execução das aplicações implementadas e a Fig 3(b) demonstra o *speedup* obtido.

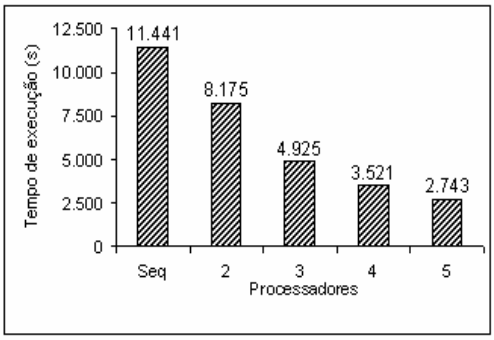

Fig 3(a). Tempos de Execução

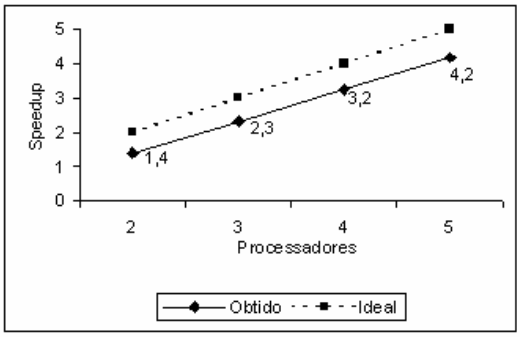

Fig 3(b). Speedup obtido

A redução do tempo de processamento ocasionou uma aceleração no processo de pós-mineração. Esta aceleração pode ser visualizada na Fig 3(b) onde é comparada com a aceleração ideal das aplicações que utilizam mais de um processador.

Conclusões

Este trabalho apresentou uma implementação paralela do processo de pós-mineração de regras de associação obtidos a partir de dados que registram o acesso à páginas da *Web*. A implementação do modelo proposto e sua execução em um agregado de computadores proporcionaram sua validação. Os algoritmos que compõem o modelo mostraram-se eficientes em relação ao objetivo de reduzir o tempo de processamento. Com sua utilização, o analista dos resultados poderá realizar um maior número de consultas a uma mesma base de regras num determinado período.

Referências

[BRU 00] BRUSSO, M. J. **Access Miner**: uma proposta para a extração de regras de associação aplicada à mineração do uso da Web. Porto Alegre: PPGC da UFRGS. Dissertação de mestrado. 2000. 95p.

[BRU 01] BRUSSO, M. J.; NAVAUX,, P. O. A.; GEYER, C. F. R. **Um modelo para a Mineração de Regras de Associação Aplicado ao Uso da Web**. In: Encontro Nacional de Inteligência Artificial, 2001, Fortaleza. Anais do XXI Congresso da Sociedade Brasileira de Computação. Fortaleza: SBC, 2001. v.1.

[FAY 96] FAYYAD, U.M. *et al.* **From Data Mining to Knowledge Discovery: An Overview** In: Advances in Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996. 611p. p.11-34.

[FEL 98] FELDENS, M.A.; MORAES, R. L., PAVAN, A. **Towards a Methodology for the Discovery of Useful Knowledge Combining Data Mining, Data Warehousing and Visualization**, In: Conferência Latino-americana de Informática, 1998, Memórias. Quito. Pontifícia Universidade Católica Del Equador: v. 2. p. 935 – 947.