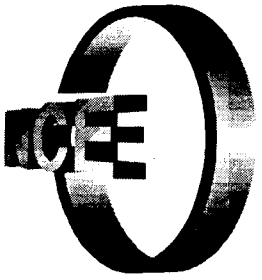

Relatório Técnico

Núcleo de Computação Eletrônica

Uma Proposta de Avaliação da Arquitetura Exokernel para Mineração de Dados

Renata da Silva Camargo
Antônio Carlos Gay Thomé
Verônica Lagrange Moutinho dos Reis

NCE - 11/99

Universidade Federal do Rio de Janeiro

Proceedings of IV Brazilian Conference on Neural Networks - IV Congresso Brasileiro de Redes Neurais
pp. 888-999, July 20-22, 1999 - ITA, São José dos Campos - SP - Brazil

Uma Proposta de Avaliação da Arquitetura Exokernel para Mineração de Dados

Renata da Silva Camargo[1], Antônio Carlos Gay Thomé[2], Verônica Lagrange Moutinho dos Reis[3]
[1]IM-NCE/UFRJ
[2]Instituto de Matemática – IM/UFRJ
[3]Núcleo de Computação Eletrônica - NCE/UFRJ
Caixa Postal 2324, CEP 20001-970
Rio de Janeiro, RJ, Brasil.
E-mails: {rcamargo, thome, veronica}@nce.ufrj.br

Abstract

This paper has been originated from the need to optimize data mining applications and to reduce their computacional resource demand. Recent researches are tipically concentrated on the search for faster algorithms. Another approach, presented in this paper, intends to provide the operating systems with facilities to better support those applications workload.

In order to validate this approach, we propose the implementation of customized libraries in the Exokernel architecture.

1. Introdução

No cenário atual em que as empresas possuem grandes quantidades de dados, armazenados em meios magnéticos, a questão fundamental é: já que eles existem, o que fazer com eles? Este é um questionamento comum em diversas organizações, onde a competitividade mercadológica e a busca por maiores faturamentos são obtidos por meio de um melhor gerenciamento da informação.

Uma tarefa importante é separar as informações irrelevantes das relevantes, com vistas a melhorar o processo da tomada de decisão.

A existência destas grandes bases de dados corporativas e a necessidade de reorganizar os dados de forma útil, tem propiciado o surgimento de técnicas e ferramentas computacionais utilizadas na busca e extração de conhecimentos.

Tais técnicas, conhecidas genericamente como *Knowledge Discovery Databases – KDD*, possuem ferramentas eficientes para o processamento e a descoberta de informações importantes, porém nem sempre óbvias, a partir de grandes coleções de dados, visando o auxílio no suporte à decisão. Por meio destas técnicas é possível analisar, avaliar a situação atual de uma organização e prover meios de acesso aos dados relevantes ao processo de decisão. Com isso, pode-se obter uma visão útil dos dados, transformando-os em informações relevantes, resultando num processo decisório mais rápido e eficaz. [17]

Os benefícios obtidos por este novo paradigma são, dentre outros: predições mais corretas, baseada em análises de tendências; economia gerada pela melhor qualidade das informações; melhor estratégia de marketing nas atividades fim da organização; melhor conhecimento dos relacionamentos entre produtos e serviços da organização.[3]

A maioria dos trabalhos de pesquisa publicados [1, 7, 10, 12, 14], que tratam da descoberta de conhecimento, concentram-se em otimizar soluções de *software* em nível da própria aplicação, preocupando-se muito pouco com as características de *software* básico ou de *hardware*, responsáveis pelo suporte à execução dos algoritmos destas aplicações.

Os algoritmos de *KDD* se caracterizam pela complexidade dos mecanismos de busca, que demandam grande capacidade computacional. Neste contexto, a pura otimização do *software* aplicativo muitas vezes não produz os resultados esperados. Em se tratando de otimização, observa-se que grande parte das limitações está geralmente relacionada com a não utilização das facilidades e recursos oferecidos pelo sistema operacional.

Diversos pesquisadores têm demonstrado que os sistemas operacionais tradicionais são inflexíveis e incapazes de fornecerem a funcionalidade e o desempenho que as aplicações necessitam [2, 6]. Uma solução seria utilizar um sistema operacional configurado especificamente para o problema. Proporcionando às aplicações não-privilegiadas um maior acesso às facilidades e recursos do sistema.

O *Exokernel* [6] é um exemplo de arquitetura de sistema operacional, concebida com o objetivo de oferecer às aplicações executadas em modo usuário, o controle eficiente dos recursos de *hardware* e *software* do sistema.

O principal objetivo deste trabalho é investigar a viabilidade da criação de bibliotecas específicas para aplicações de mineração de dados, utilizando uma arquitetura de sistema operacional do tipo *Exokernel*.

Este artigo está dividido em seis sessões: na primeira é feita uma rápida introdução; na segunda descreve-se o processo de *KDD*; na terceira descreve-se a fase de mineração de dados; na quarta apresenta-se os conceitos

principais da arquitetura *Exokernel*; na quinta discute-se a utilização da arquitetura *Exokernel* em suporte a uma aplicação de mineração de dados e, finalmente, na sexta e última sessão são feitas algumas conclusões.

2. Knowledge Discovery Database

Com o rápido crescimento da tecnologia da informação, constata-se que a sociedade moderna apresenta uma forte tendência de gerar e coletar dados em volume e velocidade cada vez maiores, resultando em grandes quantidades de dados tipicamente armazenados em meios magnéticos.

Para consolidar suas informações, as organizações defrontam-se com um novo desafio: classificar grandes quantidades de dados, no menor tempo e utilzando menos recursos, com intuito de acelerar o seu processo de tomada de decisão.

Para estruturar e recuperar essas informações, as organizações precisam de sistemas que forneção: acesso eficiente a grandes quantidades de dados, mecanismos para analisar dados e métodos de avaliação de alternativas de negócios. Este contexto é atendido pelas técnicas inclusas em Knowledge Discovery Database – *KDD* [5, 7].

KDD consiste em um conjunto de técnicas e ferramentas computacionais capazes de extrair e descobrir conhecimentos não óbvios, a partir da manipulação de diferentes bases de dados, visando principalmente o auxílio no suporte à decisão[5].

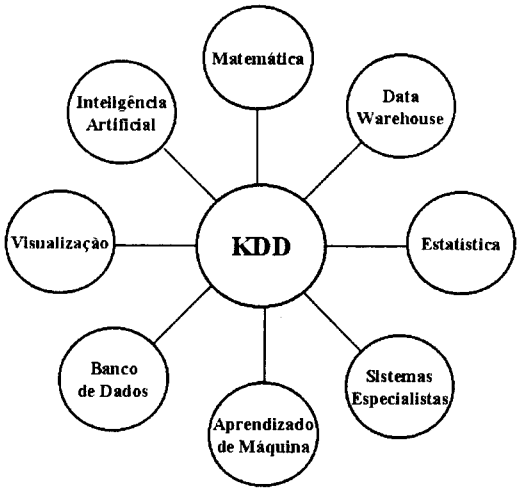

Figura 1 – Campos envolvidos no processo de *KDD*

O processo de *KDD* envolve vários campos de estudos, tais como (figura 1): Estatística, Tecnologia de Banco de Dados, Aprendizado de Máquina, Sistemas Especialistas, Inteligência Artificial, Técnicas de Visualização e Métodos Matemáticos.

De acordo com a literatura [5, 7], o processo de *KDD* é interativo e iterativo, cuja execução obedece cinco fases distintas:

1. Seleção
   - Compreender o domínio da aplicação;
   - Buscar o conhecimento anterior relevante e os objetivos do usuário final;
   - Criar um conjunto alvo de dados em que a prospecção deverá ser efetuada.
2. Pré-processamento
   - Realizar a limpeza dos dados, eliminando ruídos, inconsistências, redundâncias, ausência de dados, erros contidos nos bancos de dados, entre outros.
3. Transformação
   - Encontrar características úteis para representar os dados, em função dos objetivos;
   - Reduzir o número efetivo de variáveis consideradas;
   - Buscar representações dos dados.
4. Mineração de Dados
   - Escolher as tarefas de mineração de dados;
   - Escolher os algoritmos de mineração de dados, selecionando métodos para uso na busca de padrões;
   - Utilizar o algoritmo de mineração de dados, buscando padrões de interesse.
5. Interpretação/Avaliação
   - Interpretar os resultados e, caso necessário, retornar a qualquer uma das fases anteriores;
   - Consolidar o conhecimento;
   - Descobrir e solucionar conflitos;
   - Incorporar o conhecimento.

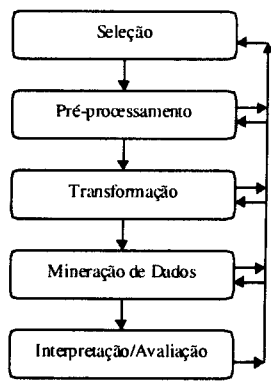

Figura 2 – Fases do processo de KDD

A figura 2 ilustra as fases do processo de *KDD*, empregado para descrever todo o processo de extração de conhecimento dos dados, enquanto que mineração de dados é uma das fases do *KDD*, conforme observado anteriormente [8, 9, 17].

3. Mineração de Dados

A mineração de dados consiste no processo de descoberta de correlações significativas, padrões e tendências, através da exploração cuidadosa e analítica de grandes quantidades de dados, utilizando tanto tecnologia de reconhecimento de padrões como técnicas matemáticas e estatísticas [7]. A motivação em se fazer mineração de dados se dá pelo enorme crescimento das bases de dados comerciais, governamentais e científicas, que vêm ultrapassando a capacidade técnica e humana na interpretação desses dados.

No contexto de uma sociedade globalizada, será cada vez mais importante a descoberta de novas oportunidades de negócio a partir da análise criteriosa do conjunto de dados armazenados pelos sistemas de informação e que muitas vezes são considerados dispensáveis após a geração dos relatórios e mapas habituais. Quanto antes essa análise ocorrer, melhor serão as condições de competitividade em relação ao mercado. As empresas que explorarem seus dados primeiro terão grandes vantagens competitivas.

Dois pontos de grande importância dentro do contexto de mineração de dados são:

→ *Descoberta de Novos Conhecimentos* – o objetivo é explicitar relacionamentos ocultos, padrões e correlações entre os diferentes tipos de dados existentes nos bancos de dados da organização;

→ *Precisão dos Dados* – o objetivo é obter dados consistentes para processamento e análise futuros.

As técnicas mais utilizadas em projetos de mineração de dados [4, 5, 7, 8, 9] incluem:

Visualização

Utilizada para analisar e observar os dados de uma determinada base de dados de forma gráfica. Ela pode ser útil para a percepção de características que se aplicam a pequenos subconjuntos dos dados, que poderiam passar despercebidas se fossem utilizados somente meios estatísticos;

Estatística

Utilizada para executar funções como agrupamento, análise e projeção de variáveis;

Regras de Indução

Processo de se obter uma hipótese a partir dos fatos já existentes, onde os fatos são os registros existentes nos bancos de dados e a hipótese é uma árvore de decisão onde os dados são segmentados de forma significativa;

Aprendizado Baseado em Instâncias

Utilizando os dados armazenados para classificar novas instâncias, baseada na similaridade de classes segundo uma determinada métrica de distância;

Aprendizado Dedutivo

Gerador de conhecimento como conseqüência lógica de um conhecimento anterior;

Redes Neurais

Utilizados em problemas do mundo real, particularmente para aqueles envolvendo identificação e reconhecimento de padrões e aproximação de funções. São modelos que "aprendem" como resolver um problema com base em séries históricas.

Algoritmos Genéticos

São métodos de busca heurística baseados nos mecanismos de seleção natural, proposto por Charles Darwin.

Estas técnicas dão suporte a um conjunto de operações; que diferem entre si pela tipologia do problema; como por exemplo [9]:

→ *Associação* – operação que retorna afinidades entre itens de uma coleção;

→ *Padrões Seqüenciais* – descobre eventos relacionados que ocorrem ao longo de um período de tempo.

As Séries Temporais, que identificam padrões sequenciais similares coletados ao longo de um período de tempo, são um caso especial desses tipo de operacao;

→ *Classificação* – classifica um dado em uma das diversas classes pré-definidas;

→ *Clusterização* – particiona o espaço do conjunto de exemplos de acordo com critério definido.

A adequabilidade do tipo de tarefa de mineração de dados ao tipo de problema, mais a qualidade e a quantidade dos dados, são fatores determinantes para o sucesso do processo. Outro fator determinante do sucesso é o custo computacional dos algoritmos de buscas e extração de conhecimentos, que demandam grande capacidade computacionais.

Diversas são as investigações desenvolvidas com o intuito de otimizar o processo de mineração de dados e os algoritmos de busca, como por exemplo os apresentados em [1, 7, 10, 12, 14].

Observada tamanha atenção dada à otimização dos algoritmos, um novo enfoque pode ser proporcionado à otimização das aplicações de mineração de dados, através da utilização de uma arquitetura de sistema operacional do tipo *Exokernel*.

4. Exokernel

O sistema operacional controla todos os recursos do computador, fornecendo a base sobre a qual os programas aplicativos executam. O sistema operacional define uma interface entre a aplicação e os recursos

físicos. Essa interface é formada por um conjunto de rotinas que oferecem serviços aos usuários do sistema e suas aplicações, bem como a outras rotinas do próprio sistema operacional. Esse conjunto de rotinas é chamado núcleo do sistema ou *kernel*. As principais funções do *kernel* são [13, 15, 16]:

→ tratamento de interrupções;

→ criação e eliminação de processos;

→ sincronização e comunicação entre processos;

→ escalonamento e controle dos processos;

→ gerência de memória;

→ gerência do sistema de arquivos;

→ operações de entrada e saída;

→ contabilização e segurança do sistema.

Uma preocupação que surge na grande maioria dos projetos de sistemas operacionais é a implementação de mecanismos de proteção ao núcleo do sistema e de acesso aos seus serviços. Além disso, os recursos precisam ser gerenciados de maneira ótima para a maioria das aplicações, ou seja, os algoritmos utilizados na gerência dos recursos do sistema visam otimizar a utilização dos mesmos como um todo. Todavia estas vantagens refletem no custo da flexibilidade e desempenho, posto que as aplicações são forçadas a utilizar abstrações e políticas específicas para acessar os recursos de *hardware*.

Por exemplo, uma aplicação poderia ter de esperar mais ou ser obrigada a usar recursos de que não necessita por ser este o procedimento padrão do sistema operacional.

Uma solução seria utilizar um sistema operacional personalizado para este problema, o que proporcionaria às aplicações em modo usuário o acesso a um número maior de instruções que são tradicionalmente executadas pelo sistema operacional.

Dentro deste principio o laboratório de ciência da computação do MIT desenvolveu o sistema operacional *Exokernel*. Tem como característica oferecer mecanismos de proteção e controle de recursos de *hardware* e *software* às aplicações. Isto propicia a concepção de ambientes personalizados de tal forma que cada aplicação enxerga um sistema operacional dedicado.

Esta descentralização do gerenciamento, no sistema operacional, é efetuada por meio de bibliotecas em modo usuário, contrastando com os sistemas operacionais convencionais, onde uma autoridade central protege e abstrai os recursos de *hardware*. O *Exokernel*, portanto, protege somente os recursos, destinando o gerenciamento às aplicações.

Na figura 3 é apresentada a estrutura básica de um sistema operacional tradicional e de um *Exokernel*:

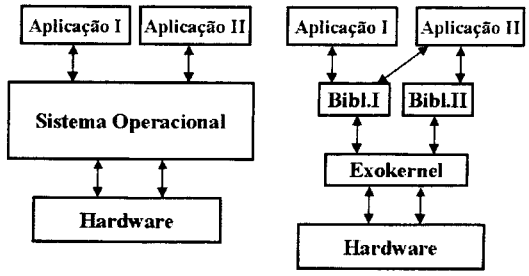

Figura 3 – Estruturas de um Sistema Operacional Tradicional e de um *Exokernel*.

Observe que as aplicações utilizam bibliotecas personalizadas, que por sua vez utilizam o *Exokernel* para alocar e desalocar os recursos de *hardware*, tais como processador, memória e dispositivos de entrada/saída. Cada aplicação, quando solicita algum serviço, realiza uma chamada a uma rotina da biblioteca personalizada, por meio de uma chamada de sistema (system call) dedicada à aplicação, com nome, parâmetros e forma de ativação específico.

Até o presente momento, diversos sistemas baseados na arquitetura *Exokernel* [11]. O Aegis, foi o primeiro sistema operacional de biblioteca, e o XOK é o mais recente *ExOS* destinado aos Personal Computer baseados nos processadores da Intel.

Diversas aplicações Unix tais como: o gcc, o perl, o apache, o tcsh e o telnet, podem ser transportadas para o ExOS. Em [11] é reportada uma avaliação, onde os resultados obtidos demostram que o desempenho destas aplicações é melhor utilizando bibliotecas específicas.

5. Um Exokernel para Mineração de Dados

Com base nas necessidade que as organizações têm em obter informações novas e úteis a partir de suas grandes bases de dados, na complexidade dos mecanismos de busca, propõe-se a utilização de uma arquitetura de sistema operacional tipo *Exokernel*, com bibliotecas dedicadas as tecnologias de mineração de dados, proporcionando assim um melhor desempenho das aplicações bem como uma maior flexibilidade, funcionalidade.

Para estudar o desempenho e a viabilidade da solução proposta, está sendo desenvolvido um projeto cuja execução está dividida em três fases:

Na primeira fase serão utilizados alguns *softwares* comerciais de mineração de dados, para observação: dos recursos mais utilizados, dos tempos de espera, do impacto das trocas de contexto, do gerenciamento de threads e outros.

Na segunda fase será desenvolvido um simulador com bibliotecas onde serão implementados os algoritmos para gerência dos recursos do sistema. Serão experimentados diversos algoritmos de forma a

determinar quais os mais adequados a uma aplicação característica.

Esse simulador será implementado em um ambiente ExOS versão 1.0.

Na terceira fase sera efetuada validação do simulador, juntamente com as bibliotecas, e a analise dos resultados obtidos.

6. Conclusões

Neste artigo foi apresentado a essência e a complexidade do processo que as organizações enfrentam para dinamizar seu processo decisório.

Apresentou-se também o processo de *KDD*, em particular mineração de dados e as técnicas disponíveis hoje para a obtenção de resultados eficientes e, possivelmente, eficazes para o suporte de decisões nas organizações.

Em face a complexidade do problema e a demanda computacional requerida, propos-se a união das técnicas de mineração de dados com uma arquitetura que suporte a utilização de bibliotecas de sistema personalizadas às aplicações.

Este trabalho está sendo desenvolvido no laboratório do Núcleo de Computação Eletrônica da UFRJ, compondo quatro projetos de final de curso de graduacao em Informatica e uma dissertaçãoo de mestrado.

Agradecimentos

Agradecemos ao Núcleo de Computação Eletrônica pelo apoio financeiro em forma de bolsa de mestrado; a equipe do laboratório da Área de Ensino e Pesquisa – AEP; ao Grupo de Redes e ao Grupo Executivo de Processamento de Alto Desempenho – GEPAD da Área de Apoio Academico – AAA. Em particular, agradecemos ao Prof. Sérgio Guedes por suas críticas e revisões.

Referências Bibliográficas

[1] Agrawal, R., Stolors, P. The Fourth International Conference on Knowledge Discovery and Data Mining. AAAI PRESS, 1998.

[2] Anderson T. The Case for Application-Specific Operating Systems. In Third Workshop on Workstation Operating Systems, 1992. p. 92-94.

[3] Berry, M. J. A., Linoff, G. Data Mining Techniques for Marketing, Sales and Customer Support. Wiley Computer Publishing, 1997.

[4] Bigus, J. P. Data Mining with Neural Networks. McGraw-Hill, 1996.

[5] Communications Of The ACM. Data Mining. November: v.39, n. 11, 1996.

[6] Engler, D. R., Kaashoek, F., O'Toole, J. Jr. Exokernel: An Operating System Architecture for Application-Level Resource Management. In Proceedings of the 15th Symposium on Operating Systems Principles, December: 1995, p. 251-266.

[7] Fayyad, U. A., Piatestsky-Shapiro, G., Smyth, P. Advances in Knowledge Discovery and Data Mining. From Data Mining to Knowledge Discovery: An Overview. AAAI Press/The MIT Press 1996, p. 1-34.

[8] Freitas, A. A., Lavington, S. H. Mining Very Large Databases With Parallel Processing. Kluwer Academic Publishers, 1998.

[9] Han J. Data Mining: Concepts and Techniques. Simon Fraser University, 1996.

[10] Heckerman, D., Mannila, H., Pregibon, D., Uthurusamy, R. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining. AAAI PRESS, 1998.

[11] Kaashoek, F., Engler, D. R., Ganger, G. R., Briceño, H. M., Hunt, R., Mazières, D., Pinckney, T., Grimm, R., Jannotti, J., Mackenzie, K. Application Performance and Flexibility on Exokernel Systems. In Proceedings of the 16th Symposium on Operating Systems Principles (SOSP), October, 1997.

[12] King, R. D., Feng, C., Sutherland, A. A.I. StatLog: Comparison of Classification Algorithms on Large Real-World Problems. May, june, 1995, p. 289-333.

[13] Machado, F. B., Maia, L. P. Arquitetura de Sistemas Operacionais. Ed. LTC, 1996.

[14] Proceedings, PADD98, The Second International Conference on The Practical Application of Knowledge Discovery and Data Mining. The Westminster Central Hall, London, UK, March, 25–27th 1998.

[15] Tanenbaum, A. S. Distributed Operating Systems. Prentice Hall, 1995.

[16] Tanenbaum, A. S. Sistemas Operacionais Modernos. Prentice Hall do Brasil, 1995.

[17] Weiss, S. M., Indurkhya, N. Predictive Data Mining a Practical Guide. Morgan Kaufmann Publishers, 1998.