Relatório Técnico

Núcleo de Computação Eletrônica

Proposing a Customized Exokernel Library to Data Mining

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

NCE - 10/99

Universidade Federal do Rio de Janeiro

Proposing a Customized Exokernel Library to Data Mining

Renata da Silva Camargo   Antônio Carlos Gay Thomé
Verônica Lagrange Moutinho dos Reis
Núcleo de Computação Eletrônica
Federal University of Rio de Janeiro (UFRJ)
Caixa Postal 2324, Rio de Janeiro,
RJ 20001-970, BRAZIL
Phone: (55)(21)598-3123
FAX: (55)(21)270-8554
{rcamargo,thome,veronica}@nce.ufrj.br

Abstract

The implementation of customized system libraries in an exokernel environment is considered as a promising approach in optimizing data mining processes. Customized libraries in exokernel environments have been successfully used in optimizing other applications, and is potentially suitable to demanding applications such as data mining. A prototype, to test our hypothesis, is under construction. This work introduces data mining, the exokernel environment and describes our prototype's building strategy.

Keywords:*Knowledge discovery database, data mining, exokernel.*

1 Introduction

Now that big companies have enormous quantities of data, stored in electronic format, a fundamental question arises: what to do with it? This is a very common question, especially because being able to better manage and use information leads to greater competitiveness and bigger earnings.

In order to better use data, it is important to separate the meaningful, relevant one, from the rest.

The existence of such large databases, combined with the need to efficiently extract useful information, paved the way to new techniques devoted to search and extract knowledge. These techniques are collectively known as Knowledge Discovery Databases (KDD). They provide efficient tools to process and discover important data, not always obvious, from very large collections of data and their goal is decision support.

The benefits incurred from this new paradigm are, among others: more accurate predictions; savings made possible from the usage of better quality data; better marketing strategies; better knowledge between products and services [3].

The KDD literature [1, 7, 10, 12, 14] is devoted to optimizing the software solution at the application level, paying no attention to the operating system or to the hardware that supports the application.

KDD algorithms incur complex and expensive search techniques. So, some optimizations may not produce significant improvements. Sometimes the limitations are strongly related to the usage of resources provided by the operating system.

On the other hand, it has been shown that traditional operating systems (OS) are inflexible and unable to provide applications with the required performance and functionality [2, 6]. One approach in trying to solve this problem is to develop customized operating systems. These OS should give flexible access to system resources to standard applications. The exokernel [6] is an example of an OS environment designed to give applications efficient control of all system's hardware and software resources, while still guaranteeing these resources' protection and security.

The main goal of this research is to determine the feasibility of developing customized libraries for KDD applications in an exokernel environment.

This paper is further divided into five sections. Section 2 describes the KDD process; section 3 describes the stages of data mining; section 4 introduces the main concepts of the exokernel paradigm; section 5 discusses the suitability of an exokernel environment in supporting data mining applications and, finally, the conclusion

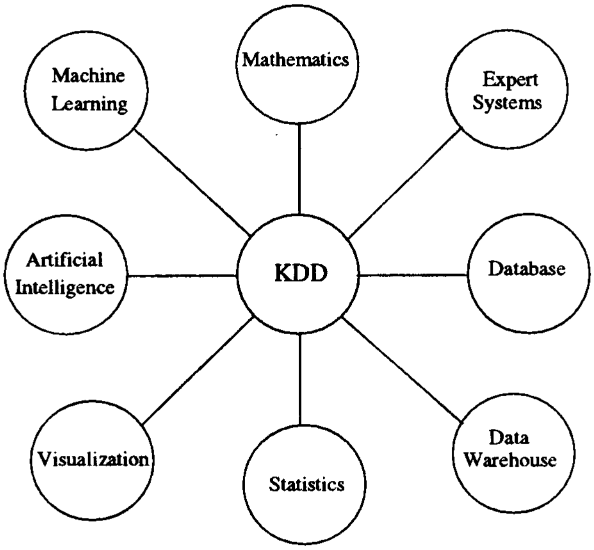

Figure 1. **Areas involved in the KDD process.**

2 Knowledge Discovery Database

As the information technology rapidly grows, a general tendency to produce and store increasing quantities of data is observed , which leads to larger data bases. The new issue is how to classify increasing quantities of data in less time, using less resources in order to increase the value associated with its decision making.

KDD comes to fill that need, as it aims at improving organizations' access to always increasing quantities of data, providing techniques to analyze and manipulate these data.

KDD is a set of tools and techniques capable of extracting and discovering knowledge that is not obvious, from the manipulation of many, very large databases, targeting decision making [5].

The KDD process uses many areas such as: statistics, database technology, machine learning, expert systems , artificial intelligence, visualization techniques and mathematics (see figure 1).

The KDD process is interactive and iterative, and follows five distinct steps [5, 7]:

1. Selection
   - to understand the application domain;
   - to search prior important knowledge and the application's goals;
   - to create a target dataset on which the search will be made.
2. Preprocessing
   - to clean the data, eliminating noise, inconsistencies, redundancies, missing data, errors, etc.

3. Transformation
   - to find useful features to represent the data, based on the desired goals;
   - to reduce the effective number of variables considered;
   - to search invariant representations for the data.
4. Data mining
   - to choose the data mining tasks;
   - to choose the data mining algorithms, the techniques to be used find data patterns;
   - to use the chosen algorithm to look for interesting patterns.
5. Interpretation/Evaluation
   - to interpret the results and, if necessary, to return to previous steps;
   - to consolidate knowledge;
   - to discover and eliminate conflicts;
   - to incorporate the new knowledge.

Figure 2 depicts the five KDD steps.

KDD is the overall process of discovering useful knowledge from data. Data mining is a particular step in this process [5, 7, 8, 9, 17]: the one that actually extracts the knowledge from the transformed data.

3 Data mining

Data mining is the process of discovering new important correlations, patterns and tendencies, through the careful exploration and analysis of very large data sets, using pattern recognition, mathematics and statistics [7]. Its main motivation is the very large commercial, governmental and scientific databases that have already exceeded the human capacity to efficiently interpret them.

From a global point-of-view, it will become increasingly important to discover new business opportunities from the careful analysis of the bulky data, stored in corporate systems and that sometimes are considered useless, after a lifetime of service generating reports and routine maps. The faster that happens, the better the competitive conditions to the organization against other organizations and the market. Organizations that first use their stored data will have a head start.

Some important goals, to data mining, are:

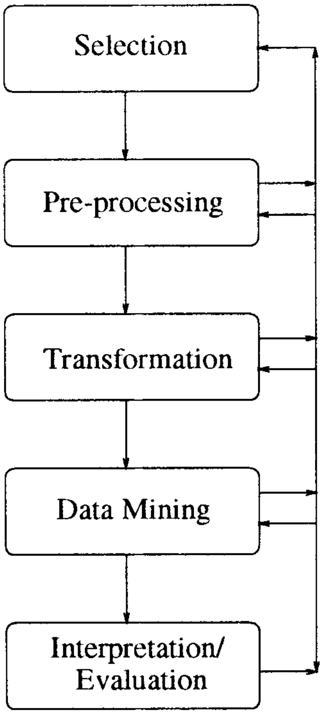

Figure 2. **Steps of the KDD process.**

- *Data Visualization*: done before any analysis. Used to qualify and organize the data to be worked on, and to discover new ways of exposing this data such that it becomes more natural and transparent to the users.
- *Knowledge Discovery*: to expose hidden relations, patterns and correlations between different types of data present in the organization's databases.
- *Data Accuracy*: to improve data consistency to future processing and analysis.

The most used techniques in data mining projects are [4, 5, 7, 8, 9]:

Visualization

To analyze and observe data in graphical form, could be useful to determine the characteristics applicable to small subsets that could not be noticed by statistics tools.

Statistics

Used to execute tasks such as grouping, regression analysis and correlation analysis.

Induction Rules

To incur a hypothesis from facts already known, where these facts are database registers and the hypothesis is a decision tree, which should group data in a meaningful way. Its main advantage is to be easy to understand.

Learning Based Instances

To use data to generate new instances, based on the approximation of classes according to some metric.

Deductive Learning

To generate knowledge as a byproduct of previous knowledge.

Neural Networks

Used to identify and recognize patterns, to approximate functions and in predictions. These models "learn" to solve a problem based on previous examples. Their "training" may be supervised or unsupervised.

Genetic Algorithms

Used to find optimal sets of parameters that describe a predictive function. These are heuristic search techniques commonly used in optimization problems.

These techniques support a set of operations and differ in the type of problems they are able to solve. For example [5, 7, 9]:

- Association: returns similarities among elements of a set;
- Sequential Analysis : models the states of the process generating the sequence or extracting and reporting deviations and trends over time;
- Classification: maps a data item into one of the several predefined classes;
- Clusterization: creates the several classes;
- Regression: maps a data item into a real-value prediction variable.

The success of the data mining process is strongly dependent of the suitability of the type of task to the problem as well as the quality and quantity of data available. Another important factor is the computational cost incurred by the complex programs used.

Currently, there are many researchers working on optimizing the data mining processes and algorithms [1, 7, 10, 12, 14]. However, we propose a new approach to the optimization of data mining applications, though the usage of a more flexible operating systems architecture. The idea is to optimize the operating system services provided to the applications by, for example, an exokernel environment using customized libraries.

4 Exokernel

The operating system controls all the resources, providing the base over which all programs execute. The operating system defines the interface between the application and the physical resources. This interface in implemented by a set of routines that offer services to users and their code, as well as to other operating system's routines. The basic set of these routines is called nucleus, or kernel. It's main functions are [13, 16]:

- to handle interruptions;
- to create and eliminate processes;
- to synchronize and communicate processes;
- scheduling;
- to manage the memory;
- to manage the file system;
- to do input/output;
- to do accounting and system safety.

One important issue, when implementing operating systems is to provide protection to the kernel. It is also important to manage resources in the best possible way to the majority of the applications. That is, the algorithms used in managing system resources must optimize their utilization as a whole. However, with these benefits, comes the lack of flexibility and, sometimes, performance, because applications are forced to use abstractions and specific policies to access hardware resources. For example, one application may need to wait more or be forced to use unnecessary resources if that's the standard procedure for this OS.

It has been shown that traditional operating systems are unable to give applications the flexibility, functionality and performance needed [2, 6]. One solution is to use an operating system that allows user applications to directly access functions originally performed by the kernel.

This is the goal of exokernel, an operating system design developed at the computing laboratory of MIT. Exokernel "exports" all management tasks to the application level and is responsible, at kernel level, to guarantee the protection of the resources by multiplexing the hardware, through the basic kernel routines

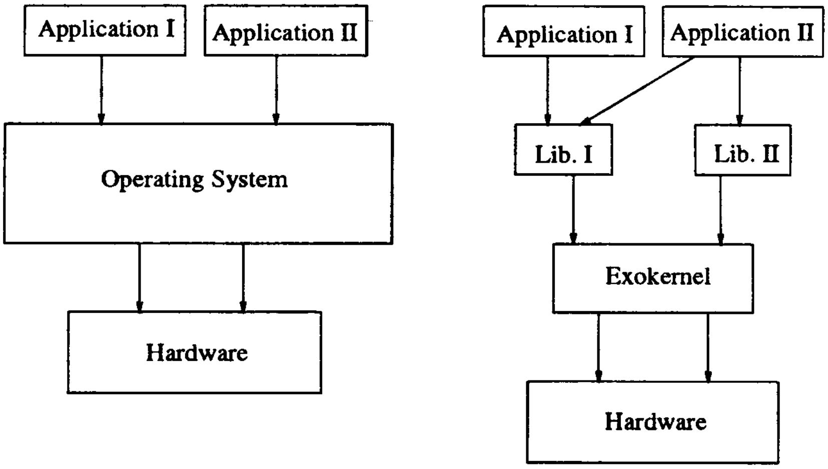

Figure 3. **A traditional operating system and an exokernel.**

and by implementing capabilities. It gives applications total control over resources, through user-level customized libraries.

Figure 3 presents the basic structure of a traditional OS and that of an exokernel. Observe that under exokernel, the applications use customized routines and that these routines will use exokernel to allocate and deallocate hardware resources such as processor, memory and I/O drivers.

When the application needs to request some service, it will call the customized library, through *system calls*. Different libraries may implement the same service in different ways, considering different needs of different applications. A previous experiment [11] showed that some applications' performance is much better when using customized libraries.

Some systems were already built, based on the exokernel design. The first one was Aegis and the XOK is the newer one executing on PCs [11]. Many Unix applications such as *gcc, perl, apache, tcsh* and *telnet* run on it.

5 Using Exokernel to Data Mining

Having in mind the urging need organizations have to efficiently obtain new and useful information from huge databases, we propose the utilization of libraries customized to data mining applications, in an exokernel environment as, for example, the XOK. Given the complexity of the search engines and their intense usage of computational power, we believe that a customized library will add flexibility, functionality and, mainly, performance to these applications.

Because most of the data mining applications limitations are strongly related to the facilities provided by the operating system, a customized library may help performance better than simple code optimizations.

The exokernel, as stated previously, was designed to provide efficient protection and control of the available resources to the user applications. This is achieved by a small kernel that exports all the hardware resources through a low level interface and using user-level libraries that provide the abstractions needed by the application [6, 11].

Observing the needs and benefits of the involved fields, came the idea of combining both, by building a customized library to serve data mining applications. The main idea is to define the appropriate policies to manage the resources used by data mining processes. To evaluate this approach, we are currently working on a project composed of three parts:

1. to use a commercial data mining software, on a predefined database. This stage's goal is to analyze and define the resource utilization, and overheads of as many data mining application as possible;

2. to develop a prototype of a library where pertaining resource management policies can be tested. For example, we will experiment with different virtual memory page replacement policies. This stage's goal is to determine which are the most adequate management strategies to data mining. This prototype will be implemented in an ExOS environment, using XOK, which is a public domain software and executed under Linux. Its current version is 1.0;

3. to evaluate and validate the prototype and libraries.

6 Conclusions

We presented the essence of the KDD process, in particular, the data mining step as well as the exokernel paradigm.

We discussed the need big organizations have to improve the access to important data and the techniques they can use. These techniques use algorithms that execute over very large databases and require very expensive computational power.

The idea of building a customized library for data mining applications was inspired by the exokernel paradigm and by the characteristics of KDD applications, that require huge computational resources.

As preliminary evaluations demonstrated the feasibility of this approach, we are currently building a prototype, at the Computing Laboratory of the Federal University of Rio de Janeiro.

References

[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., Briceo, H. M., Hunt, R., Mazires, 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.