LA-UR- 97-4725

CONF-971057--

Approved for public release; distribution is unlimited.

Title: Web Based Parallel/Distributed Medical Data Mining Using Software Agents

Author(s): Kargupta, Hillol-XCM

Stafford, Brian-XCM

Hamzaoglu, Ilker-XCM

Submitted to: American Medical Association
Fall 1997 Meeting

MASTER

Los Alamos
NATIONAL LABORATORY

Form 836 (10/96)

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible electronic image products. Images are produced from the best available original document.

Web Based Parallel/Distributed Medical Data Mining Using Software Agents

Hillol Kargupta, Brian Stafford, and **Ilker Hamzaoglu**

Computational Science Methods Group
X Division, Los Alamos National Laboratory
P.O. Box 1663, MS F645, Los Alamos, NM, 87545

This paper describes an experimental parallel/distributed data mining system PADMA (PArallel Data Mining Agents) that uses software agents for local data accessing and analysis and a web based interface for interactive data visualization. It also presents the results of applying PADMA for detecting patterns in unstructured texts of postmortem reports and laboratory test data for Hepatitis C patients.

Introduction

Data mining involves extraction, transformation, and presentation of data in useful form. As we move more and more toward a paper-less society, each of these components of data mining is likely to face the challenges of dealing with large volume of data and the very distributed nature of the data storage and computing environments.

Medical databases are often ideal candidates for large scale, possibly distributed data mining applications. In this paper we describe an experimental software agent based system for parallel/distributed data mining PADMA (PArallel Data Mining Agents). PADMA is characterized by agent based distributed data accessing, distributed data analysis, and web based interactive data visualization. This paper also presents results of applying PADMA in medical databases.

Section presents a general overview of the PADMA system. The parallel relational database accessing operations of PADMA agents are described in Section . Section describes the data analysis capabilities of the agents. Section describes the web-based user interface and visualization module of PADMA. Section and presents test results. Finally, Section concludes this paper.

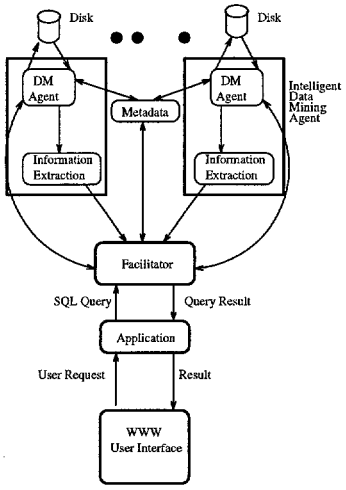

Figure 1: The PADMA architecture.

Architecture Of PADMA

The PADMA is an agent based architecture for parallel/distributed data mining. The goal of this effort is to develop a flexible system that will exploit data mining agents in parallel, for the particular application in hand. Its initial implementation used agents specializing in unstructured text document classification. PADMA agents for dealing with

numeric data are currently under development. Figure 1 shows the overall architecture of PADMA. The main structural components of PADMA are, (1) data mining agents, (2) facilitator for coordinating the agents, and (3) user interface. Each of these items are described in the following.

Data mining agents are responsible for accessing data and extracting higher level useful information from the data. A data mining agent specializes in performing some activity in the domain of interest. In the current implementation, PADMA uses data mining agents (DMAs) that specialize on both text and numeric data analysis.

Agents work in parallel and share their information through the *facilitator*. The facilitator module coordinates the agents, presents information to the user interface, and provides feedbacks to the agents from the user.

PADMA has a graphical web-based user interface for presenting information extracted by the agents to the user. The facilitator accepts queries from the user interface in standard SQL (Structured Query Language) format; the queries are broadcasted to the agents. Agents comes up with the extracted information relevant to the query. Facilitator collects the information and presents it to the user.

The agents and facilitator of PADMA are developed using a Parallel Portable File System (PPFS) (Huber, 1995). The PADMA is designed in object-oriented style to provide an extensible infrastructure and coded in C++. MPI (Message Passing Interface) is used as the message passing substrate for interprocess communication. Each data mining agent uses the underlying unix file system on the machines they are executing on for carrying out their local input/output operations. PADMA currently runs on a cluster of Sun Sparc workstations and on IBM SP-2. However it is easily portable to any distributed memory machine provided that MPI is operational on this machine and a unix file system is used for serial input/output operations on its nodes. The user interface is written for Java sensitive browser. PADMA can be functionally decomposed into three different components: (1) parallel query processing and data accessing, (2) parallel data analysis, and (3) interactive cluster/data visualization. Each of these components of PADMA is discussed in the following sections.

Parallel Data Accessing

Accessing data is an important aspect of data mining. In large scale data mining data access input/output performance becomes a critical factor in the overall performance of the data mining system. Accessing data in parallel may help decreasing the response time (Dewitt & Gray, 1992).

In PADMA, each data mining agent maintains its own disk subsystem to carry out input/output operations locally. This provides parallel data access for the whole system. Currently striped and blocked data distribution algorithms are used to distributed documents across data mining agents. Each agent and the facilitator also maintain a file cache for caching the documents that they access. Appropriate buffer management algorithms, e.g. FIFO replacement policy, write-back and prefetching, are employed to maximize the benefit obtained from these caches.

Data mining agents in PADMA also provide parallel relational database functionality. This is achieved by storing each corpus, which consists of a number of text documents, as a relational database table with different attributes. Currently a subset of SQL (Structured Query Language) is supported by PADMA. These include table creation and deletion, hash index creation and deletion, parallel select and join operations. PADMA achieves parallel query processing through intra-operator parallelism. Detailed description about the specific implementation of each of these operations can be found elsewhere (Kargupta, Hamzaoglu, & Stafford, 1996).

Parallel Data Analysis

In PADMA, data analysis is primarily done by the agents in a distributed fashion. Every agent returns a "concept graph" (which may be either a hierarchical graph of clusters, or decision trees, or statistical analysis results such as correlation matrix) to the facilitator. The facilitator is responsible for combining the concept graphs and present the result to the interface in a user transparent manner. The following part of this section briefly describes the text and numeric DMAs in PADMA.

Text DMAs

The objective of text DMAs of PADMA is to identify statistically significant document clusters that may lead to identifying common patterns among the documents in a text corpora. Text mining involves two important steps: (1) choosing/constructing the document representation and (2) finding of relations among the documents. PADMA uses a hierarchy of different representations. Relations among the documents are determined using both unsupervised hierarchical clustering algorithms and optional user feedback driven piecewise linear classifiers. Experiments reported in this paper did not use any user feedback driven supervised learning. Under no supervision state, a hierarchical clustering algorithm is used for generating a concept graph relating documents and clusters to each other. Usually clustering algorithms work from a representation of the underlying state space and a measure of similarity between any two points from the state space. Typical representation of text documents uses a vector of weighted word frequencies (Salton, Allan, Buckley, & Singhal, 1994) in the document. However, word frequency based representations are sometimes susceptible to spelling errors. PADMA text DMAs use an *n-gram* representation (Damenshek, 1995) of texts at the bottom level for rough analysis. A combination of different statistical representations and linear classifiers are used to generate a hierarchical representation of text documents (Kargupta, Hamzaoglu, & Stafford, 1996). Detailed description of the distributed implementation of this algorithm can be found elsewhere (Kargupta, Hamzaoglu, & Stafford, 1996).

Numeric DMAs

PADMA currently contains data mining agents for numeric data analysis, which are still in the developmental process. Current numeric DMAs perform simple statistical analysis such as computing correlation matrix among features and uses machine learning techniques, such as decision trees.

The following section describes the web based user interface of PADMA.

Web Based User Interface

PADMA has a world wide web based user interface as shown in Figure 2. The interface communicates with the PADMA system through a cgi script. User interface currently supports five major operations. Create option is used to create a table out of unstructured text documents. Users should supply these documents to PADMA. Read option is used to read the contents of a table. Delete option is used to delete a certain table. Query option is used to query these tables. PADMA applies the SQL query submitted by the user to the appropriate tables and presents the result back to the user. PADMA also supports analysis on a subset of the database, defined by the SQL query. Figure 2 shows the nodes of a typical hierarchical cluster among documents at a certain level. The display operation is carried out by a java applet. PADMA interface also lets the user zooming in the data and querying about specific clusters by mouse operations.

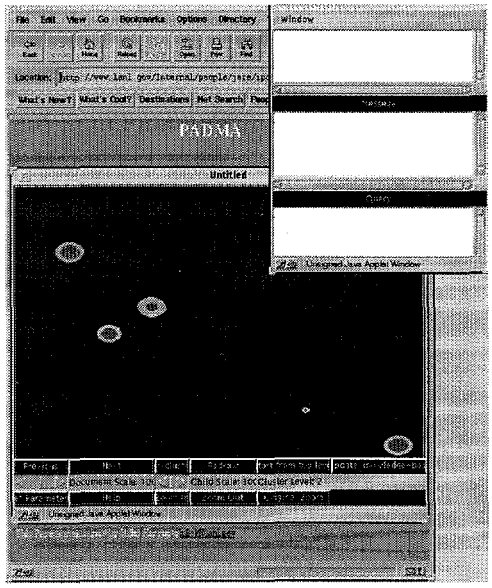

Figure 2: PADMA WWW interface contains (1) main display window and (2) windows for general query based analysis and displaying concepts/key words associated with clusters. The light-blue (gray) and red (black) colored concentric circles represent individual clusters. The diameter of the light-blue and red colored circles linearly depends on the degree of bushiness of the graph emerging from the node and the number of children leaves it has, respectively.

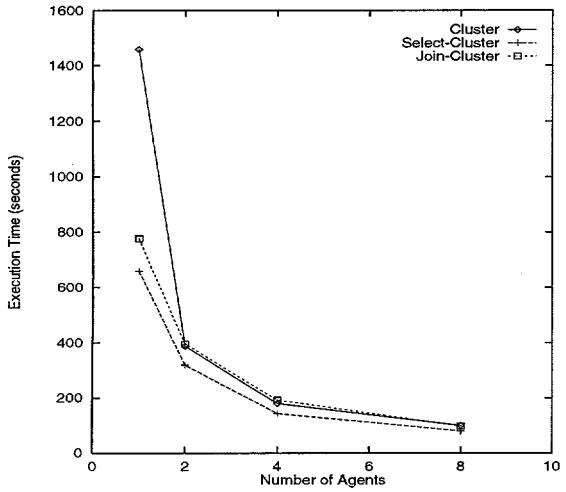

Figure 3: PADMA Performance. Note the linear speedup; initial super-linear speedup is probably due to the memory effect.

Scalability Experiments

We performed three sets of different experiments to assess the performance and scalability of the PADMA system using text DMAs only. We measured the execution times for clustering all the documents in a corpus as well as clustering a subset of the documents related through a select or join operation. Throughout the experiments PADMA agents and the facilitator are configured to use 2MB write-back caches. In all the experiments we used the TIPSTER text corpus of size 36MB containing 25273 text documents. It's striped across all agents with a striping factor of 1. The experiments are carried out on the 128 node IBM SP2 at Argonne National Laboratory. On this machine, 120 nodes are used as compute nodes and the remaining 8 nodes are used as dedicated I/O servers. Each compute node has its own I/O subsystem which uses its own local disk, and the I/O servers have faster I/O subsystems. On this machine, all PADMA components run on the compute nodes. PADMA data mining agents use the input/output subsystem of the nodes they are executing on for storing and retrieving the documents. The IBM SP2 was in multi-user mode during the experiments. Experimental results for the test of scalability are presented in Figure 3. The figure shows speedup for (1) only clustering, (2) select & cluster, and (3) join & cluster operations together. We got linear speedup for all of these cases.

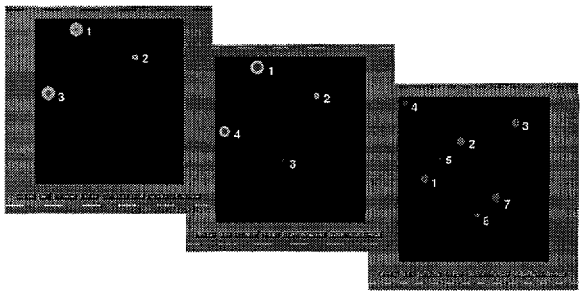

Figure 4: View of hierarchical cluster graph of postmortem report corpora at different levels.

The following section presents results of applying to medical informatics.

Medical Applications

In this section we shall briefly describe few typical applications of PADMA in medical informatics.

Postmortem Report Analysis

Postmortem reports are usually very long and prepared in unstructured text formats. Identifying death patterns from such documents is an interesting problem. PADMA was used to analyze postmortem reports provided by University of New Mexico School of Medicine. Figure 4 shows pictures of different levels of the hierarchy of clusters generated by PADMA text DMAs. As the user zoom in the data by clicking mouse buttons to specific cluster nodes shown by light blue and red concentric circles, concepts representing the nodes become more specific. For example, at the second level the clusters correspond to: (1) head and neck cases, (2) gunshot victims, (3) hemorrhage and (4) lung and heart cases. Bottom level clusters contains documents themselves. Some of the typical bottom level clusters correspond to, (1) neck injuries (2) blunt force on head, (3) blunt traumas, (4) gunshot victims, (5) hemorrhage, (6) pneumatic complications and (7) coronary artery cases.

Hepatitis-C Database

This section presents analysis of data for 545 Hepatitis C patients studied with respect to twenty two risk

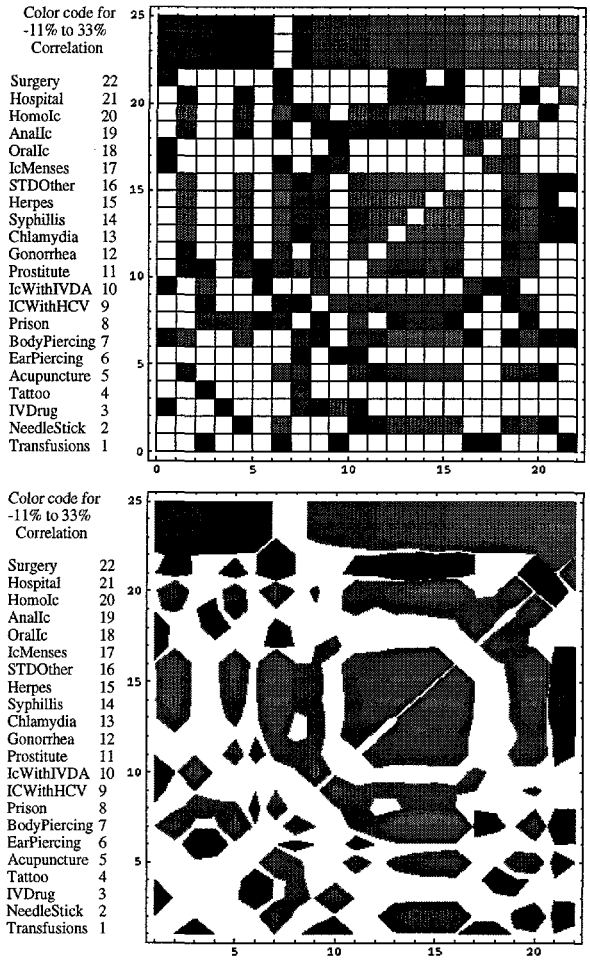

Figure 5: Density and contour plots of the correlation matrix among Hepatitis C risk factors.

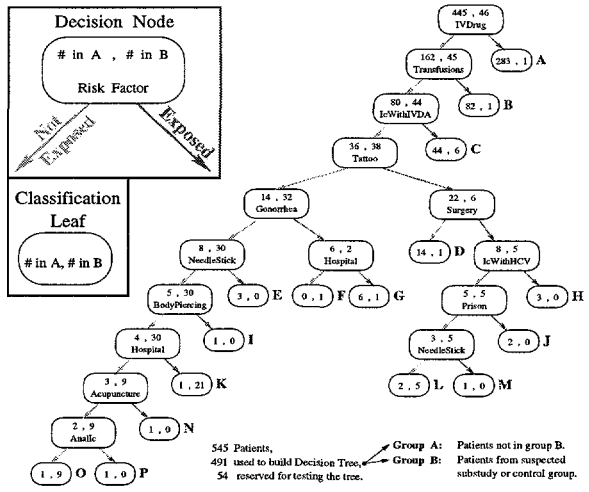

Figure 6: Decision tree model of Hepatitis C data.

factors. Distributions and correlations of risk factors are characterized for three subsets of patients (all patients, a control group (group **B**), patients with the control group excluded (group **A**)). Decision trees are introduced and used to study the contrast of the patients to the control group. The risk factors are for blood and sexual transmission of HCV (Hepatitis C Virus). Interest is in noting a separation of sexual and blood exposure which could make the data useful in characterizing sexual risk.

In Figure 5 correlation matrices are used to show how much of the variation in exposure to one risk factor can be explained by another risk factor. For example, being hospitalized is highly correlated to having surgery, and is noted at the upper left as bright red (gray). The color scale along the top shows increasing red for positive correlation, and is scaled as the percent of variation in one variable which is explained by the second variable. Negative correlations are shown in blue (dark black). For example, having been hospitalized means it is unlikely that sexual diseases were reported. The acknowledged risk factors for Hepatitis C are for blood borne transmission. Three subgroups for blood borne transmission are 1) IV drug use, 2) sexual exposure, 3) medical. IV drug use is not particularly correlated with sexual or medical exposure. Medical exposure in the form of surgery, hospitalization and transfusions are negatively correlated with sexual exposure factors. However, sexual risk is correlated to needle sticks, and acupuncture. So a pursuit of sexual risk factors must account for some blood borne risks outside of sex. One of the most interesting factors is body piercing. Body piercing is negatively correlated to most medical exposures, but is strongly correlated to sexual disease and at least positive to most of the risk factors. Studying body piercing could possibly aid understanding risks and evaluating possibilities for education or regulation to reduce Hepatitis C.

Figure 6 shows a decision tree (Quinlan, 1986) model developed from the data. The decision tree serves to classify the data into the leaves (*labeled on*

the side with bold capital letters). At each branch, a decision is made in terms of one of the risk factors. Each individual will either branch to the right if exposed to the risk factor, or to the left if not exposed. The first branch is based on IV Drug use. In leaf **A**, we see that only one (control) Group B member was exposed to risk by IV Drug use, but the majority (283 of 445) of Group A were IV Drug users. Leaf **A** classifies its members as IV Drug users. Leaf **B** classifies its members as having received transfusions, and not having used IV Drugs. A majority of Group B members which were exposed to risk factors can be seen in leaves **C,K**. Six were exposed through Intercourse with IV Drug users, and 21 through Hospitals. Also, six Group B members were exposed through Tattoos, as noted by the count of Group B members in the Surgery branch node. Leaf **O** shows that nine of Group B were not exposed to any of the ten risks along the path from the root to the leaf. The decision tree was formed to highlight the differences between Groups A and B. The distribution of risk factors must be judged in terms of providing contrast between the patients and the control group. The following section concludes this paper.

Conclusions

This paper described PADMA, an agent based architecture for parallel/distributed data mining. Main characteristics of PADMA are, (1) parallel query processing & data accessing, (2) parallel data analysis (3) web based interactive data/cluster visualization. PADMA is still under development. A module for supervised learning of piece-wise linear classifiers using feedback from the user is already developed and incorporated in PADMA. We are currently in the process of incorporating web search engines to PADMA for potential applications to web mining. We are also interested in applying PADMA for problems like disease emergence, outbreak in the near future.

Acknowledgments

This work was supported by National Center for Supercomputing Applications and US Dept. of Energy. We also acknowledge the computing time on IBM SP2, granted by Argonne National Laboratory and the data provided by University of New Mexico School of Medicine.

References

Damenshek, M. (1995). Gauging similarity via n-grams: Language-independent categorization of text. *Science*, 267.

Huber, J. (1995). *PPFS: An experimental file system for high performance parallel input/output* (Technical Report MS. Thesis). Department of Computer Science, University of Illinois at Urbana-Champaign.

Kargupta, H., Hamzaoglu, I., & Stafford, B. (1996, October). *PADMA: PArallel Data Mining Agents for scalable text classification.* Los Alamos National Laboratory Unclassified Report LAUR-96-3491. To be published in the proceedings of High Performance Computing'97.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Salton, G., Allan, J., Buckley, C., & Singhal, A. (1994). Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, *264*(3), 1421–1426.