# DataMIME™

Masum Serazi
Vasily Malakhov
Dongmei Ren

Amal Perera
Imad Rahal
Weihua Wu

Qiang Ding
Fei Pan
William Perrizo

North Dakota State University
Computer Science Department
Fargo, ND 58105, USA
{md.serazi, imad.rahal}@ndsu.nodak.edu

## 1. INTRODUCTION

DataMIME™ (Data: Mine It More Efficiently) is a 'universal' data mining system developed by the DataSURG research group at North Dakota State University. The system exploits a novel technology, the Ptree technology, for compressed vertical data representation which facilitates fast and efficient data mining over large datasets. DataMIME™ provides a suite of data mining applications over the Internet for the tasks of association rule mining, classification, and the like. The key for using the Ptree technology in DataMIME™ lies in its vertical data representation (i.e. column-based in contrast to the conventional horizontal row-based representation employed by the relational model) and processing of data which has proven the efficiency and scalability of the applications embedded in DataMIME™.

## 2. VERTICAL PARTITIONING USING THE PTREE TECHNOLOGY[1]

The concept of vertical partitioning has been studied within the context of both centralized and distributed database systems. In this system, we decompose attributes of relational tables into separate groups by bit position. In addition, we compress the vertical bit groups by using the Ptree technology.

The basic data structure exploited by the Ptree technology [1][4] is the Ptree (Predicate or Peano tree). Formally, Ptrees are tree-like data structures that store relational data in column-wise, bit-compressed format by splitting each attribute into bits (i.e. representing each attribute value by its binary equivalent), grouping together all bits in each bit position for all tuples, and representing each bit group by a Ptree. Ptrees provide a lot of information and are structured to facilitate efficient data mining processes. Once we have represented our data using P-trees, no scans of the database are required to perform the data mining task at hand. Data querying can be achieved through logical operations –such as AND, OR and NOT–referred to as the Ptree algebra in the literature. Various aspects of Ptrees including representation, querying, algebra, speed and compression have been discussed in greater details in [1][3][4].

## 3. OVERVIEW OF DataMIME™

DataMIME™ is an efficient and scalable data mining system providing the flexibility of plugging in new data mining applications when needed. Clients can interact with the DataMIME™ system over the Internet to capture their data and convert it into the Ptree format after which they can apply different data mining applications. The actual data capture along with all the data mining applications execute at the server side based on requests made by the client. Figure 1 depicts a conceptual view of the system.

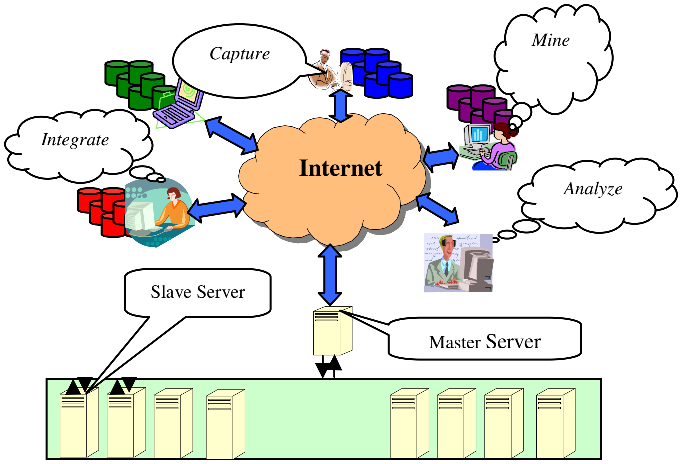

**Figure 1. The Conceptual Architecture of DataMIME™**

### 3.1 Server-side Architecture

The multi-threaded, concurrent and distributed DataMIME™ server has a four-layer architecture: DCI/DII, DMI, DPMI and DMA, and. This architecture is depicted in Figure 2.

**DCI/DII (Data Integration and Capture Interface) layer**: This layer allows users to capture and integrate data into the system. The main component of this layer is the data feeder. Tailored feeders can process particular formats of incoming data. New feeders can be integrated easily.

[1] Patents are pending for the P-tree technology. This work was partially supported by the GSA grant ACT#: K96130308

SIGMOD 2004, June 13–18, 2004, Paris, France.

**DMI (Data Mining Interface) layer:** DMI is the processing engine of DataMIME™, that facilitates the Ptree algebra, which currently has five operations: AND, OR, NOT (complement), XOR and ROOTCOUNT (the number of 1s in the bit group represented by the Ptree). These operators provide all the required aggregate processing for the application layer.

**DPMI (Distributed P-tree Management Interface) layer:** The distributed P-tree Management Interface is responsible for managing the Ptrees in a distributed environment. Note that in a distributed environment there will be more than one slave server embedded in this layer as depicted in Figure 1.

**DMA (Data Mining Algorithm) Layer:** This layer provides the suite of available data mining applications.

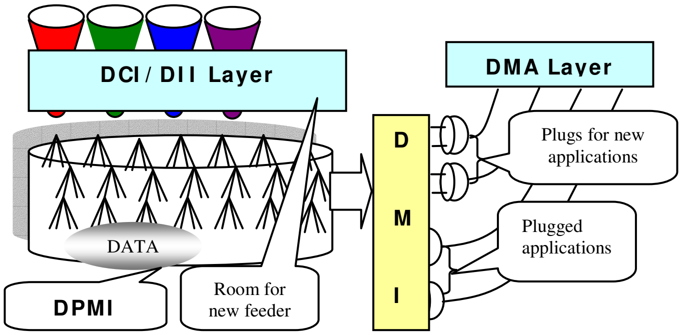

**Figure 2. Server Side Architecture of the system**

## 3.2 Client Structure

On its client side, DataMIME™ has a graphical user interface (GUI) to enable visual interaction with users. The two main functionalities of the client interface are:

- *Capturing* which sends datasets along with their meta information (description of the data) to the DII/DCI layer of the server for capturing.
- *Mining* which sends requests to the DMA layer for applying data mining applications on previously captured datasets and the presentation of the results.

## 3.3 System Features

Major features of DataMIME™ can be summarized as follows:

- The system has the ability to handle record-based, relational-data with numerical and/or categorical attributes. The data could be in text format, relational format, or TIFF-image format. In addition, capture from any other machine readable format can be easily provided through tailored feeders.
- The system provides various scalable Ptree-based data-mining applications spanning various areas such as association rule mining and classification (prediction). Once the data is in Ptree format, any generic algorithm can be applied efficiently (this is in contrast to application specific data structures tailored for the task at hand such as FP-trees [2]).
- The system has an open architecture providing high degree of software extensibility and easy integration capabilities for new Ptree-based, generic data-mining applications.
- Java-based clients of DataMIME™ can run on any platform such as Linux, or Microsoft Windows (including 95, 98, NT, 2000, and XP). The server is implemented in C++ to operate on a Linux-based platform.
- The system has a layered framework providing design flexibility. The clear separation of the processing engine from the data mining application layer can take advantage of the latest advances in hardware to provide efficient and scalable solutions. For example the server can run on a single machine or distributed across multiple machines.

# 4. DEMONSTRATION

In this demonstration we will show the process of converting a dataset into the Ptree format through the DCI/DII layer after which various data mining applications could be applied. We will show five data-mining Ptree-based applications:

- *P-kNN*: An efficient *k*NN classifier using Ptree-based distances metrics such as the HOBBIT distance metric [3].
- *PINE*: A *k*NN classifier which attempts to consider all the neighbors for a given test sample using a weighted voting mechanism [5].
- *P-Bayesian*: A Bayesian classifier showing that, by building Ptrees for the training data, we could compute the Bayesian probability values efficiently, without the Naïve assumption, thus improving the overall efficiency of the classifier.
- *P-SVM*: An SVM classifier offering a new and efficient proximal example-based approach to support vector machines with local segment hyperplane, which enable it to make optimal decisions at the local level.
- *P-ARM*: An Apriori-based approach for association rule mining that exploits the vertical data representation of Ptrees to mine frequent itemsets and association rules from datasets in an efficient manner.

# 5. REFERENCES

[1] Ding, Khan, Roy, and Perrizo, *The P-tree algebra*. Proceedings of the ACM SAC, Symposium on Applied Computing (Madrid, Spain), 2002.

[2] Han, Pei and Yin, *Mining Frequent Patterns without Candidate Generation*. Proceedings of the ACM SIGMOD, International Conference on Management of Data (Dallas, Texas), 2000

[3] Khan, Ding, and Perrizo. *K-nearest Neighbor Classification on Spatial Data Stream Using P-trees*. Proceedings of the PAKDD, Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer-Verlag. Lecture Notes in Artificial Intelligence 2336, 517-528, May 2002.

[4] Perrizo, *Peano count tree technology lab notes*. Technical Report NDSU-CS-TR-01-1, 2001, North Dakota State University, Fargo, ND, January 2003.

[5] Perrizo, Q. Ding, Denton, K. Scott, Q. Ding, and Khan, *PINE - Podium Incremental Neighbor Evaluator for Spatial Data using Ptrees*. Proceedings of the ACM SAC, Symposium on Applied Computing (Melbourne, Florida) 2003.