DATA MINING ARCHITECTURES – A COMPARATIVE STUDY

Thomas Thomas, Sanjeev Jayakumar, B.Muthukumaran.
e-mail id:shangrila81@hotmail.com
Sri Venkateswara College of Engineering
Post Bag no.3
Pennalur
Sriperumbudur 602105.
INDIA

ABSTRACT

Data mining is the process of deriving knowledge from data. The architecture of a data mining system plays a significant role in the efficiency with which data is mined. It is probably as important as the algorithms used for the mining process. CRITIKAL is a three-tier data mining architecture consisting of Client, Middle tier and the Data Warehouse. The architecture for mining semi-structured data makes a distinction between structured and unstructured data, and uses separate storage areas for them. The Kensington data mining system is an internet-based mining system for the analysis of large and distributed data sets. The Matheus *et al.*'s Multicomponent Architecture is designed to perform spatial data mining while DARWIN and PaDDMAS are used for distributed data mining. Lastly the architecture for scientific data mining, is used to mine scientific data from large science archives. The main factors that have been focused on, in these architectures are portability, scalability ,reduction in data preparation time, integration, multi-strategy and distribution .
Key Words: Comparison, Comparative study, Architectures, Integration, Data mining, E-commerce

INTRODUCTION [1]

Data Mining is the process of discovering non-obvious and potentially useful patterns in large data repositories such as warehouses.

Most organizations possess large volumes of data about their business processes and resources. While this data can provide plenty of statistical information, very little useful knowledge can be procured from it. In order to gain such useful knowledge, we need to discover patterns in the data, associated with the past behaviour of business processes. These patterns are used to dictate future strategy so as to maximize performance and profit. Such a knowledge discovery process is called Data Mining.

A Data mining architecture is a conceptual representation of the arrangement of , and interconnections between, the hardware and software components involved in the mining process. In the past few years, a number of architectural models have been developed for the purpose of data mining. Information was collected about a fair number of models, from the web. In this paper we compare nine of these models and discuss their relative strengths and shortcomings. The following is a list of the various architectures discussed in this paper:

A1. Architecture for Integrating E-commerce and Data Mining
A2. CRITIKAL Prototype Architecture
A3. Architecture for mining Semi-Structured data
A4. Integrated Data Mining Architecture
A5. Kensington Infrastructure
A6. Matheus *et al.*'s Multicomponent Architecture
A7. Architecture for Scientific Data Mining
A8. Darwin
A9. PaDDMAS

A1. Architecture for Integrating E-commerce and Data Mining [2]

Salient Features

The architecture basically consists of three components namely business data definition, customer interaction and analysis. The *Business data definition* component allows the E-commerce user to define the data and metadata (attributes) associated with the relevant business. Having a diverse set of attributes is the key aspect of this component. Apart from aiding the data mining process it also helps personalize the web experience. The *Customer interaction* component acts as an interface between the customer and the E-commerce business. A data collector is integrated as part of this component in order to analyze the various data sources like web sites, telephony and WAP. The *Analysis* component uses data transformations, mining algorithms, and OLAP tools to provide decision support. It is the abundant presence of attributes that gives this analysis component an edge over horizontal decision support systems.

The 3 components described above are connected by 3 bridges namely stage data, build data warehouse and deploy results. The *Stage data* bridge connects the business data definition and customer interaction components. This bridge transfers (or *stage*s) the data and metadata into the Customer Interaction component. In the staging process, changes can be tested before having them implemented in production. Also, replication and changes in data formats are allowed in order to improve efficiency. The *Build Data Warehouse* Bridge links the Customer Interaction component with the Analysis component. This bridge transfers the data collected within the customer interaction component to the analysis component and builds a data warehouse for analysis purposes. The build data warehouse Bridge also transfers all of the business data defined within the Business Data Definition component (which was transferred to the Customer Interaction component using the stage data bridge.). The *Deploy results* bridge transfers the analytical results such as models and scores back into the business data definition and customer interaction components. These results will be used for business rules and personalization.

Merits

- This architecture greatly reduces the data preparation time, which usually accounts for 80% of the time taken for the knowledge discovery process
- The right support is provided for the logging of data and metadata due to the emphasis on data collection at the application server layer rather than the web server layer.
- The integration between the 3 components allows the automated construction of a data warehouse within the analysis component.
- The sharing of metadata by the three components facilitates an efficient knowledge discovery process.

A2. CRITIKAL PROTOTYPE ARCHITECTURE [3]

Salient Features

It is a three-tier data mining architecture wherein the three tiers are referred to as:

- Client
- Middle tier
- Data Warehouse

Client

It has a *user interface*, which allows the user to select and submit the datasets on which data mining needs to be performed. The pre and post-processing of data is taken care of by two modules namely, *Association rule* module and *Rule induction* module. Mining for association rules basically involves, finding item sets that occur with high frequency and then generating rules based on these results. For rule induction CRITIKAL uses the concept of *Contingency tables*.

Middle tier

The frequent item sets and the contingency tables required for the processes described above are generated in this tier. Thus the pre and post processing of data involves both the client tier and the middle tier. Apart from these services CRITIKAL also provides certain algorithm independent services called *General services* which can be classified as

- Connection and access services
- Remote administration services
- Work management

The connection and access services allow the client to connect to the middle tier. Remote administration allows the middle tier software to be configured from a computer that is physically separate from the middle tier. The work manager is responsible for the execution of data mining tasks.

Data warehouse

The data warehouse tier consists of a File system and a Database Management System.

Merits

- Part of the software is implemented in Java in view of its portability and readability, thereby facilitating code development and testing.
- The prototype is able to mine very large datasets that may reside in databases or files.
- The architecture provides an optimized multi-tier support for data preparation prior to data mining thereby providing very high performance.

A3. ARCHITECTURE FOR MINING SEMI-STRUCTURED DATA [4]

Salient Features

This architecture aims to perform the mining process on semi-structured data. It makes a distinction between structured and unstructured data, and uses separate storage areas for them.

In the *Database*, every document is tagged with an *id* thereby allowing the user to easily identify the data associated with a particular document.

The *Rule generator* is the most important part of this architecture. It receives requests from the user, generates strategies based on the request and generates rules. It consists of the parser, optimizer and processor.

The *Concept library* differentiates concepts associated with different domains. It consists of concept dictionary, specialized concept guides and concept indices.

The *Information discovery* module finds and parses documents. Its main components are Discoverer, Extractor and Refresher.

The *User interface* allows the user to specify the following criteria: domain of interest, type of rule ,attribute type ,attribute value, confidence value, concept, support and background information. The rule generator generates rules based on the input provided by the user.

This architecture incorporates a compact representation of textual data that is stored in the concept library and the database.

Merits

- The distinction made between structured and unstructured data allows us to represent them more accurately, thereby generating more interesting rules.
- The concept library differentiates concepts associated with different domains. This implies that the final set of rules is more accurate.
- Since only a meaningful subset of information from a document is stored in the concept library and the database, document representation is very compact.
- **Any** database can be used as the database component. So this architecture can be built above existing databases.

A4. Integrated Data Mining Architecture [5]

Salient Features

This architecture contains the following components:

The *Data warehouse* is implemented in: Sybase, Oracle, Redbrick etc.

An *OLAP (On-Line Analytical Processing) server* enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The use of multidimensional structures allows the user to analyze the data. The Data Mining

Server is always integrated with the data warehouse as well as the OLAP server to embed ROI-focussed business analysis directly.
An advanced, *Process-centric metadata template* defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization.
The *Advanced Analysis Server* applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information decisions, whereas in other models the data is simply delivered to the end user through query and reporting software. The results hence obtained enhance the metadata in the OLAP Server by providing a dynamic metadata layer representing a filtered or abstract view of the data.
Using reporting, visualization, and other analysis tools future actions, can be planned.

Merits

- Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information.
- Integration with the data warehouse enables operational decisions to be directly implemented and tracked.
- As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions.

A5. Kensington Infrastructure [6]

Salient Features

The Kensington data mining system is used in distributed data mining. It is an internet/intranet-based system

Distribution, portability and scalability, openness, multi-user support, and persistence are the key features in Kensington.

It supports a number of pre-processing and post-processing functionalities namely: Universal data import, Visualization tools, Data analysis functions, Intuitive user interface, Visual programming interface.

The clients in the client layer communicate with servers in the server layer via the Java Remote Method Invocation (RMI) thereby enabling the system to run over the internet.

Auxiliary services form the third layer which enables the servers to exchange data with any databases which support the JDBC protocol.

Representations in PMML (predictive model markup language) and DMML (data mining markup language) are used to exchange data analysis procedures and models between servers.

Client-Side Design

It has an intuitive Graphical User Interface (GUI) implemented using Java Swing classes that supports the following:

- Interactive creation of data analysis tasks
- Viewing of sample data
- Visualization of data analysis results

The client can either reuse previously created data sets and data analysis procedures or import new data from external databases into a Table object on a Kensington server.

Server-Side Design

Novel middleware technology and a *scalable data management server* are the important aspects of the Kensington server. The middleware is implemented in Java. It takes care of all the session management tasks, thereby enabling the developer to concentrate only on the development of application business logic.

The more difficult computational tasks of data mining operations are carried out on the server, leaving the client system free from the potentially huge computational efforts.

The Kensington server consists of the following components, the implementation of which provides full performance scalability:

- Database Manager
- Table Manager
- User Space Manager
- Mining Components Manager.

Merits

- Java makes Kensington platform-independent and integrated.
- The *support for shared resources* enables data analysis to take place in groups, where data and procedures can be shared.
- The system can be customized for individual requirements.

A6. Matheus *et al.*'s Multicomponent Architecture [7]

Salient Features

This architecture consists of:

The *Knowledge base* where the background knowledge, like spatial and non-spatial data or database information is stored.

The *DB interface* fetches data from the storage or knowledge base, essentially a data warehouse that enables optimization of the queries. The use of spatial data index structures, like R-trees, improves efficiency.

The *Focusing Component* chooses the data part useful for pattern recognition. It may decide to use only some attributes that are relevant to the knowledge discovery task, or it may extract objects whose usage promises good results.

The *Pattern Extraction* module discovers the Rules in the data. Statistical, machine learning, and data mining techniques, and computational geometry algorithms are used to perform this task.

The *Evaluation* module keeps the important and significant aspects of these patterns, and eliminates the obvious and redundant knowledge.

The DB interface, Focusing Component, Pattern Extraction and Evaluation modules interact between themselves through the *Controller* part.

Merits

- The user can control every step of the mining process.

A7. ARCHITECTURE FOR SCIENTIFIC DATA MINING [10]

Salient Features

This is a 4–tier architecture developed in Java. The implementation of the client-side in Java enables users to construct queries by specifying the constraint set and the level of detail(LOD). On the server side a specific tier looks after server connections, authentication and session management. The session server constructs data requests and submits them to the data server which performs the following tasks:

- Determining the services that can fulfill the requests and submitting the requests to them
- Processing the results from these services and sending them back to the session server.

The backend consists of three components:

- A database containing the existing science archive.
- A processing subsystem that analyses scientific data and images.
- External archive and catalog interface that requests content from other astronomy sites and related resources.

Merits

- Highly interactive front end for building queries
- Ability to support third party tools using XML interface
- Provision for the user to specify set definition and LOD

A8. DARWIN [8]

Salient Features

Darwin has an open architecture that enhances its ability to implement parallelism on a wide range of platforms. It is a parallel architecture that is based on distributed-memory SPMD (single program multiple data) paradigm. *Star Data* is the main architectural component of the server. It provides all of the data manipulation and transformation operations. The toolset contains three powerful mining algorithms: *StarTree*, which uses decision, trees, *StarNet* which implements neural networks and *StarMatch* a weight-adjustable k-nearest-neighbour algorithm. Outer loop optimizers such as *StarOpt* and *StarGene* are a collection of functions for model comparison, evaluation and optimization. *StarDB* provides an ODBC link to the database. The architecture also contains a statistical analysis module, a server module and a Client/GUI module. All interprocessor communication takes place through the Message Passing Library. The processors work only on that section of the data set that is allotted to each one of them. The server supports a scalable distributed memory model of parallel processing. It is equipped with a unified data access and manipulation library.

Merits

- The distributed approach facilitates scalability in performance and portability of code. This is achieved by reducing interprocessor communications and efficiently utilizing local memory and disk resources
- The distributed memory model preserves locality, decreases collisions and improves cache utilization
- Platform-independence is made possible by minimizing the calls to platform specific features such as the operating system
- Use of C++ as the development language is beneficial as most of its compilers are platform independent.
- Availability of a wide range of data mining algorithms.

A9. PaDDMAS [9]

Salient Features

PaDDMAS is a component-based system for developing distributed data mining applications. The architecture contains a *visual program composition environment* (VPCE) which enables visual construction of an application by combining many components. These components are either local to the VPCE, or present in a remote site. The *Expert Advisor* helps users locate components of their interest. Each component is either a Java or a CORBA object, with an XML interface. The three types of components provided are:

- Data Management Components
- Data Analysis Components
- Data Visualization Components

The *Program Composition Tool* (PCT) enables a user to connect components(only if their interfaces are compatible). The *Program Analysis Tool* (PAT) ensures that data sources are connected, and that the data collecting tools are accessible. The *Program Execution Tool* (PET) generates a task graph for every application built by the PCT. The Transaction Manager resides outside the VPCE, enables recording of events, and provides continuous access to components that are used within the VPCE.

	A1	A2	A3	A4	A5	A6	A7	A8	A9
Author/Source	*[2]*	*[3]*	*[4]*	*[5]*	*[6]*	*[7]*	*[10]*	*[8]*	*[9]*
Application	E-commerce	Industrial data mining	Text mining	Business	Internet Intranet	Spatial data mining	Scientific data mining	Supervised learning	Distributed data mining applications
Java		✓			✓		✓		✓
Platform Independence		✓			✓		✓	✓	✓
Integration	✓			✓			✓	✓	✓
XML							✓	✓	✓
Multiple algorithms	✓				✓	✓		✓	✓
Distributed					✓		✓	✓	✓

Merits

- Ability to integrate third party components using XML interface
- Ability to hide component implementation, thereby enabling components to wrap .exe files in Java, C or Fortran.
- Availability of a wide range of data mining algorithms.

ANALYSIS

Throughout this paper, one word, which appears frequently, is integration. It is time and again emphasized that the data mining tools need to be integrated with the data warehouse (or the relevant data repository). Integration is said to be complete when inconsistencies are removed in both nomenclature and conflicting information, giving rise to 'clean' data. When the data mining tools operate outside of the data repository, extra steps are required such as extracting, importing and analyzing of the data. These steps that are collectively called data preparation, usually take up almost 80% of the time required for the entire data mining process. Thus, reduction of data preparation time through proper integration is of paramount importance as it directly results in higher performance.

Foremost among the limiting factors to data mining, is the size of the database. After a certain threshold size of the database the performance of the data mining tool significantly deteriorates. So, a mining architecture that can tolerate a relatively large sized database without its performance being significantly affected would certainly hold an edge over its counter-parts. Such tolerance is the main advantage of A2.

Another buzzword as far data mining is concerned is platform-independence also known as portability. In this day and age, the ability to function smoothly on a wide range of platforms is of paramount importance. Portability is achieved by choosing the appropriate development software. Today Java is almost synonymous with portability. That's why Java has been mentioned in the context of many of the architectures mentioned in this paper. Apart from portability another use of Java is readability. These two properties facilitate code development and testing. We have also seen C++ mentioned in the context of platform independence, since many of the C++ compilers are supported by most platforms.

Another tactic that has been increasingly used in recent times is multistratregy, wherein a user is given a choice of several mining algorithms. These algorithms may each represent a different approach to mining such as CART decision trees, neural networks and k-nearest-neighbour algorithms. Thus one can either choose the best algorithm for a given problem or use a combination of these algorithms to cover a larger region of the solution space. Having several algorithms can lead to the construction of more powerful and accurate models. This eclectic approach to mining algorithms will soon become commonplace.

Nowadays data mining is becoming increasingly distributed in nature. Many users at many different locations can access a set of data mining services .In this paper we have seen that A5,A9 and A8 have been developed to implement distributed data mining .Apart from multiple clients, a distributed system can also have multiple servers.

A9 is not only distributed but also component-based. These components act as providers of analysis algorithms and data management functions. The use of third party components is made possible by the use of an XML interface. This increases the scope of the datasets that can be mined, since more analysis tools are made available. The implementation of these components can also be hidden. In the course of time, we'll probably see a marked increase in the number of component-based architectures.

In traditional data mining we generally deal with datasets, which are integer types and have a modest dimensionality, whereas when it comes to scientific data, significant amount of processing is needed on floating point values. Hence A7 is designed specifically, to deal with such sparse datasets with high dimensionality.

CONCLUSION

An attempt was made to compare the relative merits and demerits of the above listed architectures the analysis of which is given in the previous section. It was found that no architecture was able to meet all the requirements of data analysis. Based on this, the following suggestions are given:

- Standardization of architectures aiming towards a uniform architecture for data warehousing
- Standardization of querying techniques used in data mining
- Standardization of the parameters involved in distribution of data mining services, possibly the establishment of an exchange server
- An XML interface to make third party integration possible.
- Tool box showing the list of available mining algorithms
- Use of platform independent development software

An international forum must be set up so as to implement the above suggestions effectively.

Acknowledgement

We sincerely thank our Management, Secretary, Principal and our HOD for all the support extended by them in our efforts to complete this paper.

References

[1] Dr Akeel Al-Attar , MD of Attar Software: Data Mining- Beyond Algorithms. http://www.attar.com/

[2] Suhail Ansari, Ron Kohavi, Llew Mason and Zijian Zheng: Integrating E-commerce and Data Mining:Architecture and Challenges. *WEBKDD'2000 workshop on Web Mining for E-Commerce -- Challenges and Opportunities, Aug 2000.*

[3] Ralph Rantzau and Holger Schwarz. A Multi-Tier Architecture for High Performance Data Mining.

[4] Lisa Singh, Bin Chen, Rebecca Haight, Peter Scheuermann, Kiyoko Aoki. A Robust System Architecture for Mining Semi-structured Data. *The Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98).*

[5] Krzysztof Koperski ,Jiawei Han, Junas Adhikary {koperski, han, adhikary}@cs.sfu.ca School of Computing Science Simon Fraser University Burnaby, B.C., Canada V5A 1S6 http://db.cs.sfu.ca/GeoMiner/survey/html/node5.html.

[6] Kensington infrastructure for E-business and E-science knowledge discovery http://www.inforsense.com/solutions/kensington.html.

[7] An Integrated Data Mining Architecture. http://iproject.online.fr/intro1.html

[8] P. Tamayo, J. Berlin, N. Dayanand, G. Drescher, D.R. Mani and C.Wang .Darwin: A Scalable Integrated System for Data Mining. *Data Mining and Knowledge Discovery.*

[9] Omer Rana, David Walker, Maozhen li, Steven Lynden, and Mike Ward. PaDDMAS: Parallel and Distributed Data Mining Application Suite. *Proceedings of the Fourteenth International Parallel and Distributed Processing Symposium, held 1-5 May in Cancun,Mexico*.

[10] P.Dowler, D.Schade, D.Durand, S.Gaudet. Scientific Data Mining.*Astronomical Data Analysis Software and Systems IX ASP Conference Series, Vol. 216, 2000.*