EVALUATION CRITERIA FOR DATA MINING SYSTEMS

ABSTRACT

Abstract- Fulfilling quality attributes of a software system relies upon developing a well-suited architecture. To
develop such architecture, one needs to know its definition as well as the quality attributes it needs to pursue.
Nowadays, data mining systems are widely used to discover knowledge from large distributed repositories. These
systems require specific quality attributes different from traditional software systems. Consequently, their
architectures differ from the traditional software systems’ architectures due to including specific elements to
fulfill their specific requirements. In this paper, we investigate data mining systems, their structures, and the
quality attributes they need to meet. Our main contribution is to present a set of criteria to evaluate the
architecture of those systems. In addition, we present a table comparing them based upon the criteria presented.
Finally, we propose a parallel architecture for distributed data mining systems which meets most of the criteria
required in these systems.

KEYWORDS

Software Architecture, Quality Attributes, Data Mining Systems, Parallelism, Repository.

1. INTRODUCTION

Architecture is a blueprint of the system that is an abstraction to help manage the complexity of the system. It
comprises software components, external visible properties of them, and their relationships to reason about
the design [Garland 03] [Kazman 02]. As the complexity of the system increases, its decomposition into
subsystems and their inter-connections (its architecture) so as to fulfill the requirements (functional or
nonfunctional) become vitally important. A good architecture has the following attributes: it is complete, if
every requirement and every system issue have been addressed. It is comprehensible by various stockholders.
Additionally, it is easily extensible. Also, it is consistent, if it does not contain any contradictions. It needs to
be realistic, so as to be possible to be implemented. Its structure needs to be highly cohesive, that means the
number of dependencies within a subsystem should be high. It is desired to be loosely coupled, that means
the number of dependencies between subsystems should be minimum [Garland 03] [Bruegge 04]. Furthermore,
due to the demands to represent architecture to various stakeholders, here we aimed at the logical view of the
architecture introduced in [Kruchten 95].

In this paper, we focus on data mining systems due to their wide use in analyzing and discovering
interesting pieces of information and knowledge from a huge amount of data stored in databases, data
warehouses, or other information repositories. So far, many data mining techniques and systems have been
implemented. They are responsible for analyzing, cleaning, integrating, selecting, transforming, mining,
evaluating, and presenting data. As a result, they are required to do the following activities: classification,
estimation, prediction, and association rule discovery [Han 06]. To pursue these functions, they contain a few
key elements such as data, mining engines, user interfaces, and concept hierarchies. On the other hand, they
are required to fulfill some nonfunctional requirements which we will discuss in details in section 2. Thus, an
appropriate data mining system architecture to meet those requirements is vitally important.

In this paper, at first, we investigate the nonfunctional requirements of data mining systems in section 2.
Moreover, in this section we present criteria to compare these systems’ architectures. In section 3, a data
mining system architecture is introduced, which we refer to it as the reference architecture. Section 4 looks
chronologically at the work done to implement data mining systems and analyses their architectures
comparing them in a table based upon the criteria mentioned. Regarding the quality attributes and the
discussed criteria we propose an ideal architecture to pursue the requirements in section 5. Finally, section 6
concludes the paper with a brief summary and future work.
2. ISSUES ON DATA MINING SYSTEM ARCHITECTURE

Data mining systems differ from traditional software systems due to their specified nonfunctional requirements. The nonfunctional requirements are as follows [Sodiya 06][Raghavan 03][ Yaginuma 00][ Zaki 02][ Maniatty 00][ Zhang 07]: Extensibility, the ability to extend the system due to having large amounts of data sources. Data Integrity, ensuring data as a whole regarding distribution of data in heterogeneous sources. Data transparency, the system should be able to cope with data heterogeneity and complexity. Distribution and data ownership, data are owned by multiple entities distributed geographically, which make us consider the number of distributed calculations, and data security. Efficiency, the system should be able to optimize the distributed data mining process to provide the users with the best response time possible. This is mostly fulfilled by predicting users’ behavior based on their previous behavior. Flexibility, the system should be generalized and adjustable to the users’ requirements. It should support multiple data mining engines, so that the user can select the most appropriate one according to the characteristics of the data to be analyzed and the way the extracted information to be used. Reliability, the system should be able to withstand from component or environmental failures by avoiding wrong manual data generation, and default values. Accuracy, this refers to the correction of the system’s functions. Simplicity, it refers to the simple structure of the system and the following aspects fulfill it: adaptability to traditional systems, a less number of sent messages, a less number of transferred data, and a less number of communications. Privacy-preserving, mining data without violating their privacy, this is achieved through establishing protocols for data collection, inference control, and information sharing. System/Data transparency, the system should be able to seamlessly access file systems, databases, or data archives. Customizability, due to the large use of these systems in business, both data and system need to be customizable. To achieve these requirements, data mining systems require specific elements putting well together.

We choose these criteria to evaluate data mining system architectures: correctness, comprehensibility, extensibility, integrity, efficiency, supporting high dimensional data, flexibility, privacy preserving, customizability, usability, transparency, supporting large amounts of data, and parallelability.

So far, many data mining systems have been proposed. In the following section, we introduce a reference architecture for these systems. We refer to this architecture as the reference architecture of data mining systems, because it includes the required elements to mine data.

3. A REFERENCE DATA MINING SYSTEM ARCHITECTURE

The architecture of a data mining system may have the following major components [Han 06]: Database, data warehouse, World Wide Web, or other information repositories, on which data cleaning and data integration techniques may be performed. Database or data warehouse server, which is responsible for fetching the relevant data, based on the user’s data mining request. Knowledge base, which is the domain knowledge to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Data mining engine, that is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis. Pattern evaluation module, that typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns. User interface communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

Data mining systems have been widely used in a variety of applications, such as business, artificial intelligence, medicine, and etc. Alternative data mining systems may add or remove elements to the reference architecture to suit specific requirements.
architecture based upon their needs. As a result of that, many applications with different purposes and consequently different architectures have been developed, which we discuss in the next section.

4. WORKS ON DATA MINING SYSTEM ARCHITECTURE

Table 4.1 compares data mining systems’ architectures regarding the fulfillment of criteria mentioned in section 2. The first row chronologically presents several data mining systems and the first column presents the criteria previously discussed.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>DM Systems</th>
<th>High performance data mining system</th>
<th>The Quest</th>
<th>DBMiner</th>
<th>CRITICAL</th>
<th>DMS</th>
<th>DATAFRONT/Server</th>
<th>UMiner</th>
<th>A component-based architecture</th>
<th>Anteater</th>
<th>WekaG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctness</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Comprehensibility</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extensibility</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integrity</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficiency</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supporting High Dimensional Data</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flexibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Privacy Preserving</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customizability</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usability</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transparency</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supporting Large Amounts of Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parallelability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The stars on each cell show the fulfillment of the criteria. The major functionality of these systems is to discover appropriate sets of interesting patterns. The systems mentioned above implement a wide spectrum of data mining functions, thus the relevant cells for correctness have been given stars. Comprehensibility has been given a star if the architecture is easy to comprehend due to its simple structure. If the users are not involved with system/data complexities in a system, transparency is given a star. In the following paragraphs we chronologically discuss the systems’ functionality as well as the criteria they pursue, which are shown in the table.

**High performance data mining system:** The system is highly generalized so as to be customized to various users’ needs. Moreover, it supports multiple independent mining engines so that the user is able to select or add the ones they need. This could lead the system to be flexible and extensible. The client/server architecture of the system has a well-defined GUI and visualization tools that clients find it easy to interact. Due to appropriate APIs, adding any data repositories seems easy, which makes the system scalable to support large amounts of data. Other criteria are not considered as the main features of this system [Yaginuma 00].

**The Quest data mining system:** The client/server architecture consists of a GUI in the client machine that users find it easy to interact with. It consists of high performance and non-linear scaling on very large real life databases. There is a standard stream interface defined for all accesses to input, insulating the algorithm code from data repository details, which are encapsulated in a data access API. Thus, it is easy to add new data repository types to the Quest system resulting in supporting large amounts of data [Agrawal 02].

**DBMiner:** The client/server architecture consists of a GUI which provides a usable environment with good performance due to the efficient implementation techniques have been explored using different data structures, including multiple-dimensional data cubes and generalized relations. The data mining process may utilize user or expert-defined set grouping or schema level concept hierarchies which can be specified flexibly, adjusted dynamically based on data distribution, and generated automatically for numerical attributes [Han 96].
CRITICAL (Client/server Rule Induction Technology for Industrial Knowledge Acquisition from Large Databases): According to the important pre-requisites for large scale data mining in an enterprise environment, such as: no limits on the data size, optimized performance, flexibility for different data mining techniques, support for multiple users and concurrency, full control of system resources, full control of access to data, and remote administration and maintenance, CRITICAL has been proposed having a three-tier architecture. The middle tier provides integrity. Moreover it supports connection and access services, which enable clients to connect to the middle tier and checks user’s authentication and security, remote administration services, it allows remote administrative tasks to monitor the middle tier, and work management, which receives a work specification from the client containing information on the data mining task and the required data sources, and based on that determines the data mining engine and does the data mining task [Rantzau 99].

A parallel data mining architecture for massive data sets (DMS): To work with industrialized-sized data in an interactive fashion, a parallel architecture has been proposed. The servers and data repositories could be added or removed in order to provide extensibility. The high data processing speeds attained by DMS are due to the following factors: effective parallelization of the data, efficient encoding of the data, use of simple and optimized algorithms, storing data by columns rather than row to reduce disk access time and memory requirements, using zoom-in functions in order to get rid of scanning the entire data, CPU is not short of memory, rich load balancing, and having large data sets. Additionally, this system provides usable interfaces [George 99].

DATAFRONT/Server: The system has a web-based client/server architecture. The server side of the system consists of a mining engine that is implemented as CORBA objects and works on distributed environments to provide integrity. Furthermore, simply adding data sources could make this system highly flexible and extensible. Moreover, the system provides elements like cache to provide efficiency. Users’ authentication is checked in the client machine so as to pursue privacy preserving [Ashida 99].

UMiner: It is a data mining system to improve the data analysis results. It handles uncertainty in the clustering and classification process and improves reasoning and decision making. It has a client/server architecture. The clients are authenticated to connect to the server to pursue privacy preserving. In addition, it provides visualization tools to make the result more comprehensible [Amanatidis 02].

A component-based architecture: The two layered architecture is extensible and modifiable at runtime. Typical objects of the client layer include buttons, lists and labels as well as scene graphs, which make the system easier to use. Data repositories could be easily added, thus the system is easily extensible and supports large amounts of data [Leissler 03].

Anteater: It uses a service-oriented architecture that relies on web services to solve the problems with large volume of data, computational cost of the algorithms, and diverse user population, and provides flexibility, extensibility, interoperability, and parallelism [Guedes 06]. It provides privacy preserving by authenticating the user’s validity through access elements.

WekaG: It is a Data mining grid-aware architecture (DMGA). Data mining grid allows data mining processes to be deployed in a grid environment, in which data resources and services are geographically distributed in a large number of heterogeneous incomplete data sources and records. The middle-ware could provide the system with integrity. The main advantage of DMGA is its flexibility, its use of generic and data grid services and its adaptation to data mining problems. The users are not involved with the complexities that could make the system transparent [Pérez 07].

5. A PROPOSED ARCHITECTURE

In order to pursue the maximum number of nonfunctional requirements we propose the following architecture (figure 5.1 and figure 5.2).

Our proposed architecture is a parallel client/server architecture that is an extended combination of the architectures we investigated so far. It extracts the features of them considering the tradeoffs. In the following paragraph we discuss the elements and the quality attributes our architecture could pursue.
The GUI provides the users with visualization elements, such as menus, charts, and tables to interact with. As a result, users find it easy to use the system. Moreover, Users will choose their desired engines, methods and algorithms through the GUI. Thus, the system could be customized based on the users’ needs. Multiple servers exist so as to make the data mining systems be able to extend, port, and support a large number of data resources. Additionally, through a number of servers and multiprocessors parallelism is easily possible to make the system faster. Upon receiving a request, Manager will look through its cache. If the answer is found, it will be sent to the client. Otherwise, the manager will find the most suitable Server available and send the request to them. To avoid the bottleneck made in Manager, we could connect the clients directly to the Servers which would cause high-coupling that would consequently reduce extensibility, portability, and system transparency that are desired in data mining systems.

The Server has the following architecture (figure 2).

Application Server receives the user’s request which includes choosing a mining engine, algorithms, servers, and databases. It controls the user’s authentication through Access Controller. It searches for the databases, data mining engines, and algorithms available for valid users. In addition, it sends users the results. Access Controller authenticates the users and their accessibility to resources. Concept Hierarchies provide various levels of abstraction as well as the background knowledge to be used in Pattern Evaluator to measure patterns’ interestingness. Miner does the same task as Data Mining Engine in the reference architecture in pursuing data mining functions such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis. Furthermore, optimized and more efficient data mining algorithms are possible in Miner to make the system even more efficient. Multiple data mining techniques to work with different types of data including multiple dimensional data exist in Miner. Miner ensures flexibility by providing the users with a variety of data mining engines and algorithms. They discover appropriate sets of patterns and rules. Data Convertor is responsible for data mining pre-processing. It provides the functions of sampling and normalization to prepare data to be mined.
Mining Results are used to store data mining results, so that mining could take place in a more efficient way as we discuss below.

Upon receiving a request, Application Server controls user’s authentication to access data using Access Controller, it also parses the amount of data requested. If the user is authenticated, it will check in Mining Result, in case of finding it, it will be sent to the client. Otherwise, it will mine the data to find the information needed.

Our proposed architecture provides users with the ability to scale their system through adding or removing servers. The system could handle large amounts of data and new data sources could be added. The servers could be geographically distributed; suitable modules could be added to Manager to provide integrity. Users are not involved with complex data, algorithms, and data mining activities so transparency could be met. However, in case of being authenticated they are allowed to add or remove any elements and choose their desired engines, algorithms, and repositories in order to make the system customizable. Privacy preserving is an important issue that could be met in the system by Access Controller.

6. CONCLUSION

Data mining as a process of discovering interesting knowledge from large amounts of data differ from typical systems due to their major components, and their specified functional and nonfunctional requirements. To pursue the requirements, its architecture becomes vitally important. Although, a lot of work has been done to fulfill the requirements, there is not yet a well structured and general architecture.

We investigated the nonfunctional requirements of data mining systems and the works done so far. In addition, we proposed a set of criteria to compare them. These criteria were extracted from data mining systems implemented so far. These criteria include: Correctness, Comprehensibility, Extensibility, Integrity, Efficiency, Supporting high dimensional data, Flexibility, Privacy preserving, Usability, customizability, Transparency, Supporting large amounts of data, Fault tolerance, and portability. Then we presented a table eliciting the data mining systems chronologically comparing the criteria they pursue. Finally, we proposed a parallel client/server architecture that could support most of the demands of these systems. It could provide a complete and comprehensible structure that is fast, portable, extensible, system/data transparent, secure, customizable, usable, parallel, and flexible that is able to support large amounts of data.

In the future, we intend to propose a methodology evaluating data mining systems’ architectures. We are contributing some issues to extend traditional software architecture evaluation methodologies such as ATAM to evaluate data mining systems’ architectures.

REFERENCES


