International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:3, 2010

An Application for Web Mining Systems with Services Oriented Architecture

Thiago M. R. Dias, Gray F. Moita, Paulo E. M. Almeida

***Abstract*—**Although the World Wide Web is considered the largest source of information there exists nowadays, due to its inherent dynamic characteristics, the task of finding useful and qualified information can become a very frustrating experience. This study presents a research on the information mining systems in the Web; and proposes an implementation of these systems by means of components that can be built using the technology of Web services. This implies that they can encompass features offered by a services oriented architecture (SOA) and specific components may be used by other tools, independent of platforms or programming languages. Hence, the main objective of this work is to provide an architecture to Web mining systems, divided into stages, where each step is a component that will incorporate the characteristics of SOA. The separation of these steps was designed based upon the existing literature. Interesting results were obtained and are shown here.

***Keywords*—**Web Mining, Service Oriented Architecture, Web Services.

I. Introduction

THE technologies available for information processing and analysis have allowed people to collect and store information from a wide range of sources. A few years ago reaching this amount of stored information was unimaginable. Due to this fact, several search mechanisms for the retrieval of information have been developed. Web Mining is presented as one of the technologies responsible for seeking rules and standards automatically.

Large systems are always broken down into subsystems which encompass some type of group of related services. The initial process of a project, which consists of identifying those subsystems and establishing a framework for the control and communication of subsystems, is called an architectural project and the result of this project process is a description of the software architecture [1].

According to [2], the reasons that justify the software architecture to be so important are well known. Among these reasons, architecture stands out as prompting the initial decisions of a project that will have a profound impact on all the following work of software engineering.

Over past few years of computing evolution, the distribution profile of an application, as well as the technology of computers and the software used to develop it, have changed dramatically. Several studies point out that Service Oriented Architectures (SOA) will have a great influence on the development of new systems in the very near future. According to [3], SOA will be the predominant software engineering practice, setting an end to 40 years of dominance of the monolithic software architectures.

Service oriented architecture has as its main objective the intense reuse of its components (services), so that the task of developing its application will primarily be the task of composing and coordinating the services already implemented, increasing reuse and reducing expenditure on resources.

Analyzing the characteristics of Web mining systems, it is clear that its technical architecture and technology needs to differ from conventional architectures of current systems to make them flexible to the point of integration with other tools and work by independent web platforms. The carried out research begins by identifying characteristics that SOA could provide for Web mining systems and technologies that could be used to implement these systems with the identified characteristics.

In order to meet the demand for tools that allows one to search for information of interest (in practice), and that would integrate with other tools with the use of components available on the web, a study for the design of an architecture composed of services, that would allow the recognition of the features and benefits of SOA, was performed.

Therefore, this study focuses on the identification of characteristics and advantages of the proposition of an SOA architecture implemented as services – with low coupling – allowing for external tools to interact with the services of this architecture at any stage of the Web Mining process.

II. INFORMATION RETRIEVAL

Information Retrieval (IR) is a computing area that deals with the storage of documents and the automatic recovery of the information associated with them. In this context, it is a science related to the search of information in documents and the search for the actual documents, the search for metadata which describes documents and their search in databases.

Thiago M. R. Dias is with the Federal Centre for Technological Education of Minas Gerais, Av. Amazonas, 7675, Nova Gameleira, 30510-000, Belo Horizonte, MG, Brazil; (e-mail: thiagomagela@gmail.com).

Gray F. Moita is with the Federal Centre for Technological Education of Minas Gerais, Av. Amazonas, 7675, Nova Gameleira, 30510-000, Belo Horizonte, MG, Brazil; (e-mail: gray@ppg.cefetmg.edu.br).

Paulo E. M. Almeida is with the Federal Centre for Technological Education of Minas Gerais, Av. Amazonas, 7675, Nova Gameleira, 30510-000, Belo Horizonte, MG, Brazil; (e-mail: paulo@ppg.cefetmg.edu.br).

The growing complexity of the stored objects and the large volumes of data demand more and more sophisticated retrieval processes. In the face of this, the retrieval of information presents new challenges every day.

To make real application of the large amount of available data, it is necessary to transform it into information, which, thereafter, should be analysed within the context of interest, by means of data mining tools.

A. Data Mining

Modern storage technologies available in the current database systems allow a large economy in the storage of large data volumes. In fact, storage became an (almost) endless possibility. Unfortunately, the analytical techniques for understanding the data and for the visualisation of these collections of data did not progress at the same pace.

The Data Mining tools appeared with the intention of facilitating data analysis and visualisation, as well as the discovery of useful information for decision making. They are basically computational tools that seek information in big databases – information understood as non trivial.

Data Mining is the task of extracting valid, comprehensible and effective information, not previously known, starting from great collections of useful data, in order to facilitate the decision making [4].

According to [5], Data Mining is a step in the process of Knowledge Discovery in Databases (KDD) that consists on the application of the data analysis and the discovery algorithms that produce an enumeration of particular patterns (or models) on the data.

B. Knowledge Discovery in Databases

The process of KDD was proposed around 1989 and presented a group of necessary stages to produce knowledge starting from database data. The data mining is set up in the stage of the transformation from data into information.

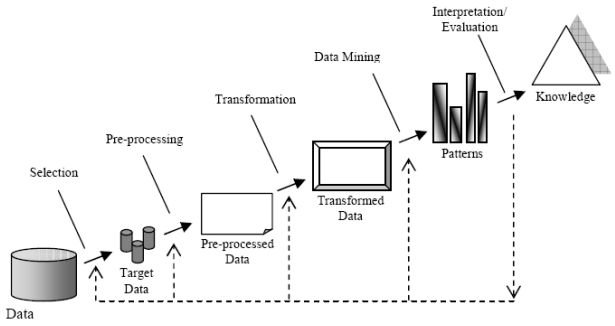

Fig. 1 Knowledge Discovery in databases [5]

In Figure 1, it can be observed that the process of knowledge discovery begins with action on the databases. This data suffers a cleaning process, where redundancies and ambiguities of data representation are removed. After data is cleaned and integrated, it is stored by subject in a Data Warehouse. Following the selection stage of which of the data should really be taken into account for the knowledge discovery, the data mining should be applied through one or more mining algorithms. With the discoveries presented by the mining, an analysis of the discovered patterns, to verify which ones really make sense for use, should be carried out.

This acquired knowledge can be seen as a systematic action that transforms intellectual goods of the organization into greater productivity, new values and competitiveness increase, to support and guide the best form of capitalising the organizational knowledge.

C. Information Retrieval from the Web

With the increase of the internet popularity, a great collection of data was generated and stored in several databases distributed throughout the world. In these banks, there is data from the most diverse areas of human knowledge.

The web also eases the publication of a diversity of material types. This allows researchers to publish the results of their scientific research in a much faster and cheaper way. Consequently, scientific literature on the Web became richer but more complex and distributed in a disorganised way over several databases. All these characteristics tend to make the internet a very difficult environment for discovering really useful material.

The process or environment in which information originated from the internet is stored, and later on manipulated and visualised, is called Web Warehousing. This environment works as a system database manager, which is responsible for storing the extracted information from the internet, especially at a data consultation and analysis level, and to establish an interface with the final user.

The internet is a vast collection of heterogeneous documents. It possesses a dynamic nature and thousands of pages appear and disappear every day. For this reason, there is concern that the web does not really reach its potential and fail to turn itself into a more usable, effective and comprehensible tool. In that perspective, the data mining appears as an obvious possibility to be explored.

The search engine services, programs designed to consult and retrieve information stored in HTML or text pages, are still very far from reaching the desired reliability level, so that only the really wanted information is retrieved. A new technique of retrieving information is currently been devised in order to overcome the difficulties of the conventional techniques now used. This technique is called Web Mining or information mining on the web.

III. WEB MINING

Over the past years, the application of Web Mining has been increasing in a dynamic environment, that is, the internet. It is growing more and more to be the standard and has been denominated as the data mining on the web. Hence, when the information mining is focused on the internet environment, the expression data mining is used as Web Mining. As stated by [6], Data Mining on the web can be defined as the discovery and analysis of useful information originated in the web itself.

Web Mining is frequently associated with the retrieval of information, but, in reality, it is treated as a wider interdisciplinary process, involving wider retrieval techniques of information statistics, artificial intelligence and data

mining. Therefore, data mining on the web is configured as a complex process of mining based on the Internet.

Web Mining is usually subdivided into three main categories that make up the interest areas from where to mine information (Figure 2):

- Web Content Mining or Mining of the Content of Documents on the Web.
- Web Log Mining (Web Usage Mining) or Mining of the Use of the Web.
- Web Structure Mining or Mining of the Structure of Documents in the Web.

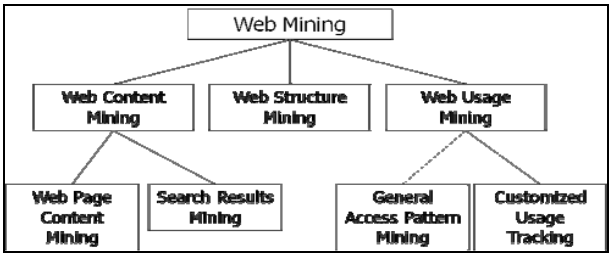

Fig. 2 Structures of Web Mining

A. Web Structure Mining

The hyperlink structures contain a vast amount of hidden information that can be the mining target. Mining in the structures of the web constitutes in the process of knowledge extraction, starting from the web documents hypertexts interconnections. According to [7], Pinski and Narin, in 1976, proposed a variation of the impact notation factor proposed in 1972 by Garfield, based on the observation that not all the citations are equally important. According to the authors, a document would have influence if, as a resource, it is highly mentioned by other influential documents.

In [8] some algorithms that find hubs and authorities are described. They are basically a page that indicates the authorities of recognized quality and authorities that are on a page pointed out by several quality hubs. It is noticed that hubs and authorities maintain a relationship of mutual reinforcement: good hubs indicate good authorities, and a good authority is that pointed out by a good hub.

Among the possible uses for Web Structure Mining are the pages placement (ranking) and the information flow.

B. Web Content Mining

A lot of the knowledge in the Web is inside documents, i.e., in their content. The discovery process of this useful information from that content is called Web Content Mining.

Current research tools are not enough aid to the user, because the search can return an amount of links that does not actually satisfy the user, or that have little to do with his/her real needs. This happens due to great part of the knowledge being located within the document, and that the search tools hardly get to reach that knowledge part for the extraction of keywords through the use of their heuristics.

Web Content Mining includes the tools that carry out the intelligent information retrieval or those that abstract the organization of the semi-structured data contained in the web. Some of those tools make use of intelligent agents while others make use of concepts based on databases. One way or another, a more efficient search or a more high-level data structuring in the internet is guaranteed.

C. Web Usage Mining

Each Web server keeps, locally, a registrations collection about the users' interaction: the access logs. Web Usage Mining uses this data to discover information on the web users, such as their behaviour and interests. If the information from the logs is well structured, it can be applied to mining techniques. The possible uses for Web Usage Mining are personalization and marketing. With that knowledge, automatic rearrangements can be suggested in the site, so that the customer is guided to pages with products that he/she would potentially buy.

As can be observed, the use of Machine Learning Algorithms and Web Data Mining techniques on data characterises Web Mining. However, the distinction between Web Content Mining, Web Structure Mining and Web Usage Mining is tenuous.

IV. SERVICE ORIENTED ARCHITECTURE

Software architecture is an abstract concept that allows a series of definitions. The definition used by ANSI/IEEE affirms that software architecture considers basically the intrinsic and extrinsic relationship among the fundamental components of a system [9]. The concept of services is the fundamental component of a SOA.

SOA can conceive a relatively cheap solution with a better cost-benefit than when referring to systems that need to talk amongst themselves and processes that demand a larger flexibility and agility to assist the market revolutions. SOA introduces a new logic (services) layer within the computations distributed platform [10].

The SOA construction begins with the definition of the application objectives. Only after the objectives have been defined, it can infer which services should be offered and how these will be contained in groups of related services.

The architecture definition, in a practical way, can be put as a development guided to services. This means that the applications will be allocated in an interdependent way – the response of an infrastructure of preset and pondered technology – to create services with enough flexibility for being reused among the systems.

Nowadays, the SOA is recognized as an important alternative for development, especially for business systems applications. It allows flexibility, as the services can be supplied both locally or outsourced.

In the present context, service can be understood as each component that assists a specific business function for their customers. A service receives the requests and answers them, hiding the entire detailing of the processing. Services can be implemented in any programming language. Also, legacy systems can be packaged as services and be used for a large range of applications without the need for large investments [1].

There are two important points in SOA: the consumer and the provider. The former consumes and requests the results to

the provider, whereas the latter executes the service and answers the needs, as depicted in Figure 3 [11].

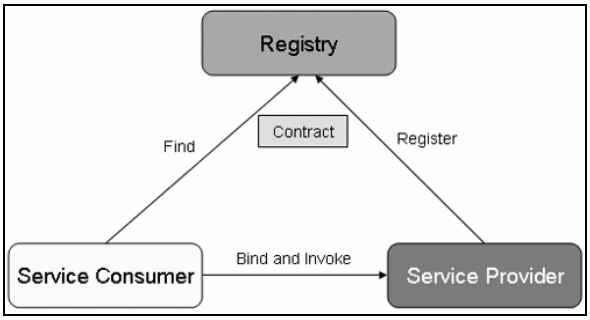

Fig. 3 Basic paradigm for SOA [11]

Every service should possess a public interface, exposed at a place where the providers can access it. To request a service, it should only be necessary to obtain the interface. Moreover, the interface should only possess the relevant functions, in a high abstraction level, taking into account the gross granularity principle. Hence, while any functionality can be transformed into a service, the challenge is to define a service interface that is at the right level of abstraction [11].

The important thing is that for the architecture to be considered SOA, the component must be a service, which have implies in: coarse-grained functionality; relevance and a high level of the executed transaction; low coupling; a very well defined interface and well encapsulated implementation details.

A. Web Services

The most important principle of SOA is its interoperability. It is fundamental that all the components communicate independently of the language in which they were built, of the operational system in which they are being executed and of the hardware architecture. Web Services conform well with the requirements since they use protocols and standard formats and accepted documents.

Today it is common to mix up Web Services with SOA. However, Web Services is a sort of incarnation of SOA, but not its definition. By its nature, Web Service technology favours the creation of components weakly coupled with gross granulation, but SOA components can be created using any other technology.

A Web Service exposes its interface for the users using a XML framework, known as Web Services Description Language, or WSDL. Using WSDL one can discover which are the data types, messages format and services made available by a Web Service.

The customers can request a Web Services using two ways: either they have direct access to WSDL or they use a registration service through an UDDI (Universal Discovery Description and Integration) interface.

The WSDL pattern defines a Web Service as a collection of net endpoints, better known as ports. A port allows for some operations and each operation implies in the exchange of some messages that are formed by defined data types in an XML outline.

V. DEVELOPMENT OF THE ARCHITECTURE

Initially, in the SOA approach implementation is a more complex task because of the need to worry about reduction of data transfer among components, in order to reduce traffic, and about services state control. These project decisions affect the construction of the system architecture, which should possess independent components. Flexibility means that system evolution process is simpler when (and if) is later performed.

The implementation of a Web Service in the context of SOA provides flexibility for the distribution of services in the system, making it easier for commitment, in contrast with other more traditional technologies. Several works show the potential of Web Service and SOA combination for data mining systems on the web [12].

A. Building the SOA Web Services

For the case study presented in this work, data were extracted from an available location, the CEFET-MG Web Site (CEFET-MG is a Federal Technical Institute in Brazil).

As a proof of concept, a Web Usage Mining Tool was developed from scratch to verify in practice the advantages and shortcomings of using a SOA framework over traditional approaches. Selection and retrieval, pre-processing, transformation, data mining and visualisation modules were implemented as services and individually tested. Each service had its WSDL document created with the support of the development platform. After its creation, the tool also allowed the creation of each customer for the service in question, in an automated manner. Only after the creation of each customer, the overall framework was tested to verify the results and implementation consistency.

After testing each individual service in place, services were grouped to verify the implementation of the whole process of Web Mining and observation of the automatic collaboration of each service with the next service.

The first service to be implemented was the selection and retrieval module. This service, which is responsible for receiving an input file from the client, has methods to process the input file and specific methods to evaluate its structure and content. If data are in the correct format, the received file is sent in an automated way to the next service that executes pre-processing.

With the help of a customer that was automatically generated by the development environment using the WSDL document, the log file is first sent to this service and the system should be checked to its structure. A fragment of this file can be seen in Figure 4.

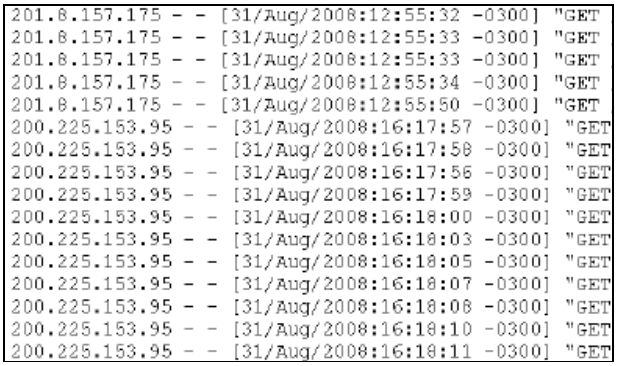

Fig. 4 Fragment of the log file

Pre-processing methods are responsible for eliminating incomplete records of the file and duplicate records, besides removal of fields and records that will not be used at the data mining stage. It also has a method that after the changes made in the file, sends the resulting file to the next service in the process of Web Mining.

The file size is an important point to be considered, since the whole process will be conducted on the web. Thus, it is necessary to eliminate as much data as possible that are not interesting and necessary for the service of mining, ie, this service also deleted data that are not useful for the mining process. One example was the exclusion of characters "[]" and ":", as can be seen in Figure 5.

What is wanted is the removal of all data that are needed or will not interfere with the results after service mining, aiming to reducing network traffic whenever is possible. The archive for this study, in its original state, had 24.2 MB. After the preprocessing and processing steps, the file sent to "Mining" had 6.05 MB, ie, the file to be mined now is almost 25% of its original size, which facilitated the traffic of this information over the internet. Thus, only data of interest will be sent to the processing service, which aims to place these data in the input to the mining algorithms.

Fig. 5 File after the step of preprocessing

The service of Transformation, which is responsible for incorporating the processing stage of the process of Web mining, consists of methods that transform the file into a format acceptable by the mining algorithm. In the specific case of this implementation, data mining module service uses the Apriori algorithm, imported from the Weka tool.

The file is transformed into a specific format for the Apriori algorithm and, thus, association rules can be extracted and data analysis have been implemented in the service of mining. The result of the file after the transformation service can be seen in Figure 6.

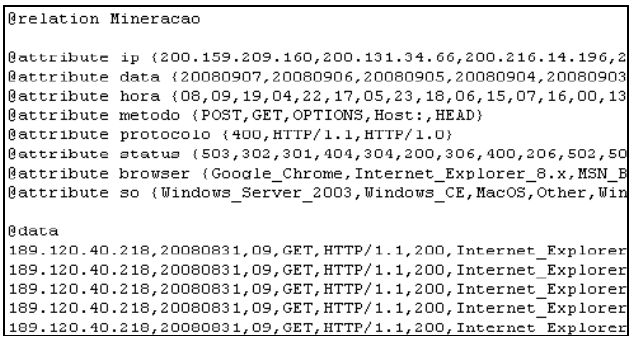

Fig. 6 Archive generated by the Service Transformation

Finally, a method sends the file already converted directly to the next service.

The data mining service has specific methods for data analysis and statistical methods to work with association. To perform statistical analysis there are methods that extract information from the file, such as number of accesses, consumed server bandwidth and access times. Parameters such as confidence and support are informed, and the package extracts association rules, returning a set of drawn rules. After these operations, a method is responsible for receiving statistical analysis and association rules for sending results to the last service in the entire process of Web Mining.

Service Mining is considered the most important service the whole process. It is exactly this service that gets information from data that, previously, were not so useful. However, all earlier services that comprise the process are considered essential to the success of mining. After performing this service, the information is sent to the Web Service View, which incorporates the latest stage of the architecture, in order to this information be placed in an easy-to-understand and structured graphics and tables.

The last service was implemented in the service view. This service is designed to receive data from the data mining service and build an HTML file and to send this file to the client that originally sent the web mining requisition. This file is stored on the server where the service is running.

To carry out the construction of the graphics has been implemented in the JFreeChart service library. This library can be used to generate pie charts, bar charts, line charts (with or without 3-D effect), Graphics, among many others. JFreeChart is written entirely in Java and can be used in any implementation of Java 2 (JDK 1.2.2 or higher). Examples of graphs generated can be seen in figures 7, 8, 9 and 10

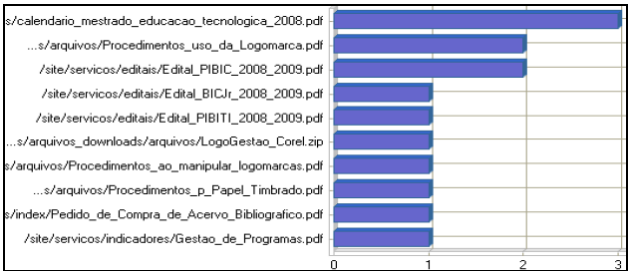

Fig. 7 Files most requested

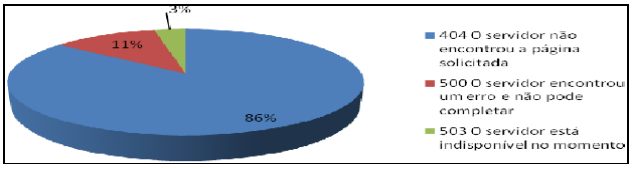

Fig. 8 Most frequent errors in the

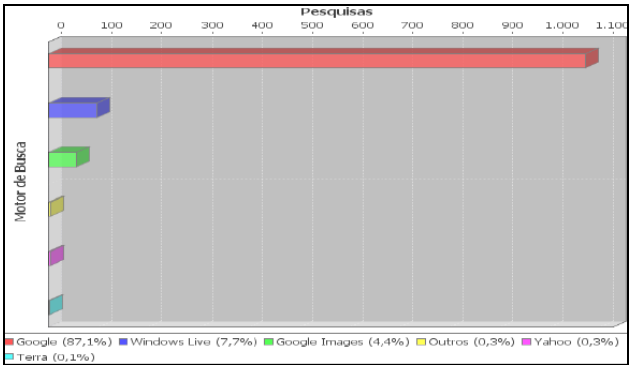

Fig. 9 Tools most frequently used search

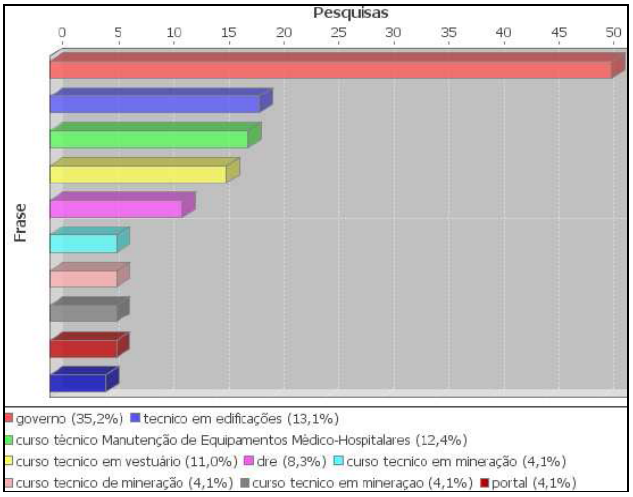

Fig. 10 Search word in the search engines

B. Performance of the Implemented Services

For this paper, the Apriori algorithm was chosen to extract association rules. The algorithm was imported from the Weka tool. Weka (Waikato Environment for Knowledge Analysis) is a system created to support the learning of algorithms. This whole system is implemented in Java and brings various algorithms used to extract knowledge in DB. It was developed at the University of Waikato, New Zealand and includes several methods for classification, association rules, clustering and prediction. It has the ability to integrate with other tools such as Web Services.

All the association rules drawn by the Apriori algorithm were analysed and considered as valid relation. They represent occurrence of values in the input file. The rules are set by combining all the possibilities of association between attributes. Finding the best rules means to select the rules with the greatest possible confidence. For this work, the rules may be useful to find patterns of behaviour of users to access a Web site.

The implementation will begin from a client that was built based on the WSDL document of the first Web service architecture that embodies the stage of selection and data recovery. This customer has the option to select a log file stored at some place and hold your shipment.

The server generates a file with the records to access the pages, and by default saved in a single file to access the data in the period of one week. Each file with a log is automatically saved to a directory on the server. The file chosen for this case study covers the period from 31/08/2008 to 09:14 am on the date to 07/09/2008 at 06:23 am, corresponding to a week. The file in question has 24.2 MB in size and consist of accesses to CEFET-MG Web Site. After sending the file to the Web Service Selection and Recovery of Data, the mining process begins.

As a final result, association rules are extracted by the implemented algorithms. After execution, the information is sent to the visualisation service, which incorporates the final stage of the architecture, putting this information in formats easy to understand, structured as charts and tables.

C. Produced Rules

As a result of the implementation of the Apriori algorithm, some association rules were obtained, as shown in Table 1.

Some rules cannot provide information that may significantly impact on business or in the structure of the site. Most of the rules listed above are called “simple”, that means, rules known by the business analysts. However, the association rules provide information that is important and useful for the Usage Mining user.

TABLE I BEST RULES FROM THE ALGORITHM APRIORI

Best Rules
Apriori
Instances: 88416
Minimum support: 0.5 (53050 instances)
Minimum confidence: 0.9
1. browser=Internet_Explorer_7.x so=Windows_XP 46236 ==> method=GET 45138 conf:(0.97)
2. browser=Internet_Explorer_7.x 45889 ==> protocolo=HTTP/1.1 44078 conf:(0.96)
3. so=Windows_XP 57479 ==> method=GET 54410 conf:(0.95)
4. so=Windows_XP 57479 ==> protocol=HTTP/1.1 53251 conf:(0.93)
5. protocol=HTTP/1.1 81155 ==> method=GET 74701 conf:(0.92)
6. status=200 71971 ==> protocol=HTTP/1.1 65894 conf:(0.92)
7. method=GET 81778 ==> protocol=HTTP/1.1 74701 conf:(0.91)
8. status=200 71971 ==> method=GET 65515 conf:(0.91)
9. method=GET status=200 65515 ==> protocol=HTTP/1.1 59600 conf:(0.91)
10. protocol=HTTP/1.1 status=200 65894 ==> method=GET 59600 conf:(0.90)

For further details on the implementation and experiments, refer to [13].

VI. FURTHER CONSIDERATIONS

The use of an SOA has proved an interesting solution for creating a platform for Web Mining. As stated earlier, although this assessment is not complete, we can conclude that this model has many advantages when compared to the traditional model where all components of the systems are embedded in a single program and characteristics of SOA described previously could not be obtained as easily. The use of SOA forces the developer to think differently, since there is no single place to access the information and program data. This, at first, becomes more complex to implement; it is important to worry about the reduction of data transfer between components to save network traffic. These design decisions will greatly affect the construction of the system architecture, which must consist of independent components. This flexibility, however, makes the process of expanding the system much simpler.

To try to make the process more agile, it was performed a compression of SOAP envelopes before sending to the next service. The service receives this envelope, performs the unpacking of these before you start processing and, thus, the process continues until the last service, with the display of results. Compression, before the data is sent from one service to another, occurs at all stages of the process.

After these changes, new tests were performed in order to verify the gain with the bundles. Working with SOAP envelopes for each service compressed, the whole process took an average of 310 seconds, which were once the 450 seconds. It is observed that there was a considerable reduction in run-time process, compared to the time before compression. In addition there was an analysis to check which service architecture was being spent a greater amount of time in execution.

After verification of each service that incorporates a phase of architecture, it was noticed that the first service architecture consumes on average about 70% of the total processing time. The justification for this high percentage in this service, and also the justification for the large increase in processing time between running the web and on a local network, is given by the fact that they have to send the log file on the Internet. Due to file size, the process of sending the file spends much of the time of execution. Another service that draws attention to analyze the time spent on executions was the mining service. This service consumes an average 15% of the total, mainly because of the demand for association rules.

To perform the averaging 10 executions of services on a computer connected to the Internet running with 01 Pentium IV processor with 4 GB of memory were carried out. Based on what was described above, one can say that the performance of the whole process will be much better if these services are performed within a local network, where it is not necessary to relies on the internet and there is no need to send the entire file on the web, process that consumes much of the time spent on implementation. Importantly, all executions were performed with the same log file that contains the area around the site of CEFET-MG, that is, all tests were performed with the same file.

A recurring question usually it concerned with the maximum number of requests that the architecture implemented can meet simultaneously. You should be aware that each request will consume a requested amount of memory, then the memory is a limiting factor for the maximum amount of requests. The major difficulty is that at the time of the request, processes that are not tied to Web services can run on the server where the services are stored and thus alter the measurement of how a play is consuming.

It is noteworthy that this is the first implementation of the architecture given only one sub-area of Web mining, specifically mining of Use But for our problem, the architecture showed positive results. The architecture developed was implemented with the concepts of SOA and the results of its implementation have brought some consistent information. Finally, we obtained a functional platform Mining Web-based services. With this work, we could demonstrate the great potential of using SOA and Web Services.

Another advantage that can be considered is the ability to easily extend the architecture to meet other sub-areas of Web Mining, with the definition of new services, to also reach the Mining Structure and Content Mining. The impact on the implementation and availability of these new features would be much greater if it were used SOA.

An important feature is the ability to send a query to any Web service, regardless of the stage where it is found must always follow the WSDL description of the service order that meets the characteristics of this service, and result in processing from the service requested.

With this implementation the features and advantages described above may be absorbed by the systems and providing the ability to integrate with other external tools.

VII. CONCLUSION

The data mining in the internet environment comes as an application and research area with excellent future perspectives due to the variety of available information in it.

There are some advantages to be obtained with the application of SOA as they were presented. The application of these advantages in Web Mining systems can provides several benefits in the information retrieval from the web, bringing quality to both the usage mining, as well as the content mining or in structure mining, seeking excellence in results and serving the dynamic and heterogeneous nature of the Web.

To create services that will assist the functional objectives of the system is enough for the success of the architecture. In addition, it is very important to provide services with the desired quality to ensure, for example, an acceptable response time and scalability of the system.

In this work, it was demonstrated the great potential of using SOA and Web Services. Another major advantage that can be considered is the ability to easily expand the architecture to meet the other sub-field of Web Mining.

Having developed the SOA tool and after the experiments, it is easy to note in practice the benefits SOA framework can provide during the debugging, evolution and integration tasks.

REFERENCES

[1] Sommerville I., "Software Engineering", 8th ed., Pearson Addison Wesley, 2007.
[2] Pressman R.S., "Software Engineering", 6th ed., McGraw-Hill, 2006.
[3] McCoy D.W., V.N. Yefim, "Service-Oriented Architecture: Mainstream Straight Ahead", 2003.
[4] Simoudis E., "Reality Check for Data Mining", IEEE Expert, 11(5), 26-33, 1996.
[5] Fayyad U.M., "Data Mining and Knowledge Discovery: Making Sense out of Data", IEEE Expert, 11(5), 20-25, 1996.
[6] Cook D.J., L.B. Holder, "Graph-Based Data Mining", IEEE Intelligent Systems, 15(2), 2000.
[7] Zaiane O.R., "*WEB* Mining: Concepts, Practices and Research", in "Proceedings of the Brazilian Symposium of Databases", Tutorial, XV SBBD, João Pessoa, Brazil, 2000.
[8] Slattery S., T. Mitchell, "Discovering Test Set Regularities in Relational Domains", in "Proceedings of the 17th International Conf. on Machine Learning", Morgan Kaufmann, San Francisco, 2000.
[9] ANSI/IEEE, "Recommended Practice for Architectural Description of Software-Intensive Systems", ANSI/IEEE Std 147, 2000.
[10] Erl T., "Service-Oriented Architecture: A Field Guide to Integrating XML and Web Services", Pearson Education Inc., 2004.
[11] Mahmoud Q.H., "Service-Oriented Architecture (SOA) and Web Services: The Road to Enterprise Application Integration (EAI)", Sun Microsystems Inc, 2005. Retrieved 04 April 2008 from java.sun.com/developer/technicalArticles/ WebServices/soa.
[12] Guedes D.O., W. Meira Jr., R.A.C. Ferreira, "Anteater: A Service-Oriented Architecture for High-Performance Data Mining", IEEE Internet Computing, 10, 36-43, 2006.
[13] Dias T.M.R., An Service Oriented Architecture for Use in Web Mining, M.Sc. Dissertation, CEFET-MG, Brazil, 2008. (In Portuguese.)