A service oriented architecture to provide data mining services for non-expert data miners

Marta Zorrilla *, Diego García-Saiz

Department of Mathematics, Statistics and Computation, University of Cantabria, Avda. Los Castros s/n, Santander, 39005, Spain

Keywords:
Analytics service
BI-as-a-Service
Knowledge discovery database
Data mining
Service-oriented architecture
Web Services

A B S T R A C T

In today's competitive market, companies need to use discovery knowledge techniques to make better, more informed decisions. But these techniques are out of the reach of most users as the knowledge discovery process requires an incredible amount of expertise. Additionally, business intelligence vendors are moving their systems to the cloud in order to provide services which offer companies cost-savings, better performance and faster access to new applications. This work joins both facets. It describes a data mining service addressed to non-expert data miners which can be delivered as Software-as-a-Service. Its main advantage is that by simply indicating where the data file is, the service itself is able to perform all the process.

1. Introduction

In a market as competitive and global as today's, currently affected by a deep economic crisis, information is one of the main managerial assets since its analysis helps in effective steering, as De Leeuw [35] pointed out 28 years ago.

Regardless of the size of the company, the need for having an accurate and reliable knowledge of what is affecting its business and for discovering new useful information hidden in the data for correct decision making has meant that since the end of nineties, business intelligence (BI) tools have been used more and more although the sector growth has not been so high in the last few years as a consequence of the economic crisis [72].

Business intelligence tools, as is well-known, encompass a wide range of techniques and technologies which are used to gather, provide access to and analyze data from the operational systems of the organization and other external sources (for instance surveys, information from competitors or data from the web, among others) with the aim of offering decision makers a more comprehensive knowledge of the factors affecting their business and, in this way, help them to take more accurate and effective managerial actions.

Among the different elements which make up a BI environment [33], we consider four of them, the data warehouse (DW), the On-Line Analytical Processing (OLAP) technology, the reporting tools and the data mining techniques to be the most essential.

The DW is the integrated repository of the strategic information of the organization which generally includes measurements, metrics and facts from the different business processes of the company (known as key performance indicators — KPI). These measurements are defined according to the different users' perspectives. The OLAP technology meets managers' and business analysts' needs to quickly search and explore accurate, up-to-date, complete information from the DW, this information being detailed as well as aggregated. The reporting tools and, in particular, dashboards and scorecards aim to help analysts to monitor and analyze the status of their KPI and drill into detailed data to identify the root causes of problems and intervene while there is still time. Lastly, the data mining techniques facilitate the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and models which can be used directly in decision making (for instance, a model for preventing credit risk).

Nowadays the majority of large companies and corporations have to a greater or lesser extent a DW and they use reporting and OLAP tools to extract and analyze the information which allows them to position themselves strategically in the market. However, although there are areas where data mining techniques are being used more and more, such as business [48], marketing [61], education [16], banking [46], health systems [78], and so on [52], their use is still not generalized. This is mainly due to the fact that data mining projects need highly qualified professionals (expert data miners) to achieve, in reasonable time, useful results for business. According to Fayyad et al. [20], these results must be non-trivial, valid, novel, potentially useful, and ultimately understandable patterns to be able to be used in decision making. One of the reasons for which expert data miners are required is that the knowledge discovery in databases (KDD) process involves multiple stages [20], and regretfully, in each one, there is a large number of decisions that have to be taken with little or no formal guidance. The lack of a theoretical framework that unifies different data mining tasks [77] explains why the KDD process is said to be as much an “art” as it is “science” [45,71].

Except for some specific cases, business intelligence needs can be grouped in domain specific solutions as for example retail banking [27], insurance risk assessment [63], discovering web access patterns [34,81], selective marketing campaigns [8,71], acquiring and retaining customers [26,32], and so on. Since the information which companies have in their transactional systems as well as the questions they want answered have a lot in common, generic data mining models can be designed in order to satisfy the needs of all of them. One easy way to define these models is by means of templates, which specify the data set to be processed, the kind of result which is required (for instance a segmentation, a rule set or a predictive model), the pre-processing tasks to be carried out and the mining algorithms to be used. These templates would be defined by a data miner, expert in the business domain, and exploited by all the users who access the service proposed in this work.

As far as we know, there is no service in the cloud which allows an end-user to extract patterns and models by simply sending his data file without having to carry out the tedious job of selecting attributes, pre-processing and setting data mining algorithms. A service like this does not only offer non-expert data miners a tool for analysis but also facilitates the work of the expert data miners who can use it to obtain initial patterns easily and quickly.

In short, our objective in this paper is to describe a software architecture which meets the necessity of non-expert data miners to extract useful and novel knowledge using data mining techniques in order to obtain patterns which can be used in their business decision making process. Our proposal follows a service-oriented architecture with the aim of being easily configured and hosted in the web and can be deployed as an Analytic Software-as-a-Service. Furthermore, a service-oriented architecture implemented by means of Web Services facilitates its extension with new functionalities (services), developed by ourselves or by third-parties (through an orchestration of services). Another additional advantage that SOA offers is its design, based on layers, which allows the improvement of certain parts of the system without affecting the rest.

This paper is organized as follows. First, we write a preliminary section in which our interpretation of some concepts and terms used in the paper are explained. Next, we review the context of BI-as-a-Service and enumerate some currently available on-demand tools. Likewise, we relate other works published with a specific focus on the knowledge discovery process and discuss these in relation to our proposal. After that, we describe the architecture of our service and discuss some details about its implementation. In Section 4, we present an application which uses the proposed data mining service, called E-learning Web Miner, which allows virtual course instructors to extract knowledge from the clickstream stored in the e-learning platform logs. And, finally, we close by summarizing the contents of this chapter and discussing our future work.

2. Preliminaries

In the last few years, a set of terms and concepts have appeared which are not clearly and accurately defined in the software world and all of them are used profusely. We refer to terms such as business intelligence as-a-Service (BIaaS), Analytics as-a-Service, Software On-demand, business intelligence in the Cloud, service-oriented architecture (SOA), Web Service (WS) or service-oriented computing (SOC) among others. In this section, we do not attempt to define these terms but indicate the sense in which we understand and use them in this work.

When we talk about a data mining service we understand "service" as a software product which offers a solution or gives an answer to the needs of a customer, this being either a person or another software application. So, there are at least two parties involved, the service provider and the service consumer, although a third party could exist, a service broker, which would act as the intermediary. Here we link with the On-demand term which means, in our view, the ability for customers to have instant access to the service and pay for it based on usage, only if this service is not free. In general, these services are offered across the Internet and therefore, On-demand Software and Software as-a-Service (SaaS) are used as synonyms. According to the Software & Information Industry Association [65], SaaS applications are based on a recurring subscription fee and typically follow a pay-as-you-go model. However, according to Srinivasa [66], currently most SaaS are free, as for example, web applications for communication and collaboration offered by Google or recently Office Web Apps offered by Microsoft.

SaaS applications are characterized by being: easy to use, feature-rich, easy to access and they promise good consumer adaptation. Generally, SaaS is used to refer to business software rather than consumer software since this delivery model avoids the need to install and run the applications on the computer of the user and to carry out the maintenance and support tasks. So, the adaptation of the SaaS concept to provide business intelligence services is known as business intelligence-as-a-Service (BIaaS).

Another relevant characteristic of SaaS applications is that they run entirely in Cloud Computing which, according to NIST [41], is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. That means, Cloud Computing provides environments to enable resource sharing in terms of scalable infrastructures (storage, computing power, etc.), middleware (databases, operation systems, application servers), application development platforms, and value-added business applications which can be delivered through a SaaS model or as utilities, namely loosely coupled sub-processes inside customers' business processes.

Much like other software, SaaS can also take advantage of service oriented architecture (SOA) to allow software applications to communicate with each other. Each software service can act as a service provider, exposing its functionality to other applications via public brokers, and can also act as a service requester, incorporating data and functionality from other services.

Channabasavaiah et al. [7] defines SOA as follows: "SOA is an application architecture within which all application logic is defined as services, which can be called in defined sequences to form business processes". Erikson et al. [18] gather seven different definitions of SOA which come from organizations such as W3C, IBM, or OMG among others and conclude that SOA is commonly seen as a way of assembling, building or composing the information infrastructure of a business or organization. Additionally, SOA Manifesto [64] states that SOA is a type of architecture that results from applying service orientation, a new way of conceiving and designing the software, focused mainly on the business processes of an organization, known as Service-Oriented Computation (SOC). Unlike previous architectures, SOA focuses on business processes, rather than technical components.

Although an official set of service-orientation principles does not exist [19,44], there is a common set that is mostly associated with service orientation, namely loose coupling, autonomy, reusability, statelessness, abstraction, composability, and discoverability. In spite of the fact that there is a variety of technologies and standards to implement an SOA including RPC, DCOM, CORBA or WCF, Web Services are the most widely used [50]. One of the main reasons is that Web Services are based on open standards that are independent from any implementation platform. Some of these standards are: XML (eXtensible Markup Language) for writing data exchange files and messages which are used for the communication between the service requester and the service provider; XSD (XML Schema Definition) for providing a means of defining the structure, content, and semantics of XML documents; SOAP (Simple Object Access Protocol) for transporting these messages in an envelope across the Internet [13]; WSDL (Web Services Description Language) for describing the details of the service such as the functionality it offers, how it communicates, and where it is accessible; UDDI (Universal Description, Discovery, and Integration) for registering the Web Service in a service broker; WS-BPEL (Web Services Business Process Execution Language) [14] for modeling business processes in terms of composition of Web Services; WSS (Web Services Security) for providing security to Web Services and WSN (Web Services Notification) which allows event driven programming between Web Services and provides support for publishing/subscription mechanisms [50].

3. Related works

Traditionally, business intelligence applications have been architected with a focus on the back-end, which is generally supported by a data warehouse. This meant that companies had to invest a lot of money in software which would allow them to build the DW and explore and analyze the information stored in it. This was only feasible for large companies and organizations. Therefore, the suppliers of BI tools now provide small and average-sized companies the possibility of moving their systems to the cloud with the aim of saving costs, getting better performance and having rapid access to new applications. This means companies are consumers of BI services hosted in servers in the cloud which support the scalability required and use grid-based system hardware.

Currently, almost all BI tool vendors have announced a strategy for the cloud deployed as a PaaS (Platform as-a-service) or as a SaaS model. For example, GoodData [22] offers a complete platform for storing and analyzing data. SAP BusinessObjects, the world's leading business intelligence software company, provides a hosted on-demand platform [59] to deliver analytic and reporting functionality. PivotLink [53] offers a SaaS solution which covers data analysis, reporting and dashboards. MicroStrategy [42] offers its BI platform for hosted reporting, analysis and monitoring software. LogiXML [38] provides a fully web-based data integration environment with ad-hoc reporting, analysis, dashboard and data integration applications. Panorama Software's PowerApps [49] allows users to query hosted data by offering its web-based OLAP engine to the general public. RightScale offers RightScale-BI [55], a cloud based business intelligence service based around open source products from JasperSoft, to create a complete analysis solution.

In relation to commercial data mining software, Adapa [1] is the first real-time scoring engine available on the market and accessible on the Amazon Cloud as a service. LityxIQ [36] is an integrated suite of analytic solutions for non-technical marketers hosted in a cloud computing environment which includes modeling and optimization tools. IBM offers its Smart Analytics System [62] which comprises data mining and unstructured information analytics. In the non-commercial field, we find Biocep-R project [11], an open source platform for the virtualization of Scientific Computing Environments (SCEs) such as R and Scilab which can be run on high performance machines or on the cloud. It enables geographically distributed collaborators to view and analyze terabytes of data interactively and collaboratively. There are other non-commercial general purpose data mining tools such as Anteater [23], GridMiner [5] or ADaM [58] which provide data mining services for expert data miners to construct KDD processes as a composition of services that are available over infrastructures distributed on a large scale. All of them offer their own graphical user interface for designing and executing the KDD process. There are other initiatives in the same line, such as [60,69] which also provide data pre-processing functions, data mining algorithms and visualization techniques, wrapping these, as we do, by means of Web Services, but none of them wraps the full process which is the goal of our tool. In our proposal, the service offers data mining models and patterns for specific problems, previously defined and studied, without the need for the end-user to have data mining knowledge. That means, the service establishes the definition of the attributes which the data set must have, the pre-processing tasks to be carried out and the selection of algorithms and their settings in order to answer a specific business problem, whereas the end-user (human being or machine) only needs to indicate where the data is when the service is requested. The service carries out all the knowledge discovery process by itself.

Although most of the mentioned data mining tools have been built according to a service-oriented architecture, none of them is currently offered as a service on-demand.

As mentioned previously, the KDD process from data involves the repeated application of several steps according to Fayyad and Piatetsky-Shapiro [20] which can be summarized in:

1. Developing an understanding of the application domain and the goals of the end-user.
2. Creating a target data set: selecting a data set or focusing on a subset of variables, to which the discovery process is to be applied. This step in general requires one or more of the following tasks: data cleaning and pre-processing, removal of noise or outliers, data transformation, adding context information, etc.
3. Choosing the data mining task, this means, deciding whether the goal of the KDD process is descriptive or predictive, and based on this decision, choosing the data mining algorithm(s) and the most appropriate parameters for its execution.
4. Executing the algorithm on the data set.
5. Interpreting mined patterns.
6. Consolidating discovered knowledge.

It must be noted that as a consequence of the existing need in the market for a systematic approximation to the development of data mining projects, companies and software consulting firms have designed process models to guide the user in this task. These process models gather the same phases proposed by Fayyad et al. but use language more orientated to the end-user. The most used models are SEMMA (Sample, Explores, Modify, Model, Asses) proposed by SAS and CRISP-DM (Cross-Industry Standard Process for Dates Mining) proposed by a consortium of European companies in the sector.

As can be observed in the previous enumeration, the KDD process requires the data miner to make many decisions in each step of the process (selecting variables, choosing data mining algorithms, setting parameters for the algorithms, etc.) but, when the data set and the domain of the business are well-known, a general way of functioning which obtains correct and useful results for making decisions can be defined. Our service works in this way. It offers several templates which contain the definition of the attributes (data) as well as the pre-processing tasks, the mining algorithms and the parameter-setting which are adequate for obtaining the patterns as explained in Section 5.

This idea of utilizing templates was used in [30] to build a single unified environment that data analysts could use for carrying out KDD projects based on a similar project which was stored in a library. Its goal was to assist analysts to do their work easily and quickly based on the reuse of other projects. Kietz et al. [31] utilize this same idea to define workflow templates which help data miners to correctly connect different tasks of a KDD process and check its correctness before its execution. It is based on an ontology which encodes rules from the KDD domain on how to solve DM tasks. In [21], the authors present a template model to help users define the multidimensional inter-transactional associations to be mined and, in this way, speed up the discovery process.

As far as we are concerned there is no service similar to our proposal although the underlying idea is that which companies typically use to give response to several customers of the same business domain. What do exist are many agents or software applications which use data mining techniques internally to offer services of clustering [24,9], classification [24,37], personalization, prediction [57], search, etc. in very different business domains using transactional data as well as data from the web [10,68] as input data for the KDD process. But, in general, they are offered by means of a user interface, except Internet searchers like Yahoo [76] or Bing [3] which are also offered by means of Web Services.

4. Architecture of the data mining service

The term SOA describes a concept for aligning an enterprise's IT environment with its business process. This is achieved by providing loosely coupled atomic services that can be flexibly combined with one another. An SOA can be implemented with the help of any arbitrary service-based architecture, but Web Services are most commonly used [50].

Before starting with the architecture description, we must say that our service has been designed as a complete service, functioning autonomously, this means, it does not require any other component or service to work, although in the future, it could be orchestrated with other services in order to offer a more powerful functionality. For this reason our service has been designed following the SOA principles and implemented by means of Web Services.

In order to explain the service architecture using a reference framework, we used that proposed by Arsanjani [2]. Fig. 1 depicts an adaptation of Arsanjani's architecture for our service.

The architecture of our service is divided in five layers. Data layer (first layer) gathers the Data Mining Service Repository and other data sources which store data to be processed by the service. The data access is based on a wrapper which mediates between calls from client application components to the data sources by transforming incoming requests into a message format that is understandable to the Enterprise Components.

The second layer, called Enterprise Components, gathers the components that are responsible for realizing functionality and maintaining the QoS of the exposed services in the third layer. This currently consists of four Web Services: one for wrapping data mining algorithms and the pre-processing tasks, another for visualizing the results of the obtained model, another for validating the xml data file sent to the service and transforming it to the format which the data mining algorithm requires and the fourth for connecting to and querying the repository. The communication among these Web Services is based on an XSD schema defined for this purpose as a consequence of the lack of standards for exchanging data and knowledge as Podpecan et al. stated in [54]. Although there are some advances in this direction, for instance Predictive Model Markup Language (PMML) for describing various data mining and statistical models, and ExpML language for sharing machine learning information [70], the majority are unsupported by the general community. Moreover, there is no common and generally accepted XML-based language for describing tabular and other types of data and most data mining algorithms still use old style data formats like csv, tab, or arff [74].

In our current implementation, the DM algorithms' Web Service wraps four data mining algorithms: SimpleKmeans [28], Xmeans [51] and EM [28] from Weka [74] and the implementation of Apriori (association rule miner) developed by Borgelt [4]. It presents its results in the proprietary format of the algorithm.

The WS-Visualization offers different kinds of graphs such as histograms, spider and pie charts for graphically showing clustering results and a 3D-graph for visualizing association rules [75].

All of these components have been programmed in java except for the visualization module which also uses the graphical capabilities provided by Matlab.

The third layer exposes the services which can be consumed by a client application or software which wants to include this functionality, for example, a Learning Content Management System (LCMS) as we show in our case study. This service can be discovered or be statically bound and then invoked, or possibly, choreographed into a composite service. The service is described in WSDL (see Fig. 2). A WSDL file describes four critical pieces of data: the interface information describing all publicly available functions (<definitions>), the data type information for all message requests and message responses (<types>), the binding information about the transport protocol to be used (<binding>), in this case SOAP and, the address information for locating the specified service (<service>).

Fig. 3 depicts an example of the SOAP messages exchanged between the E-learning Web Miner (ElWM) application and the Data Mining Service. The first message calls the service, indicating the template to be used and the location of the data file; and the second one, is the message which is sent by the service indicating where the result file is stored.

The fourth level corresponds to the business process composition or choreography layer. This tier is responsible for the choreography and orchestration of the services exposed in layer 3, making them act together as a single application. Languages such as BPEL4WS (Business Process Execution Language for Web Services) can be used to carry out this process. In our case, as we only offer one service,

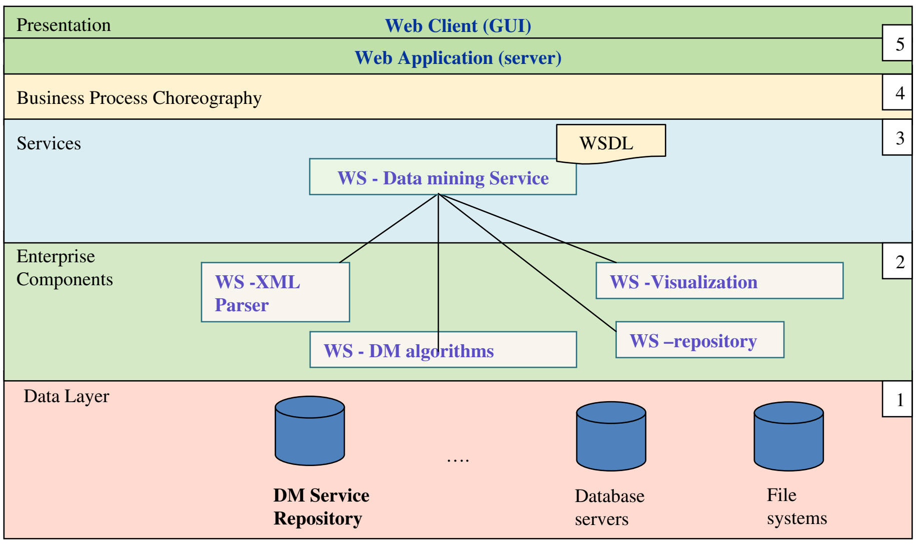

Fig. 1. Data mining service architecture.

```
- <definitions xmlns:wsu="http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-wssecurity-utility-1.0.xsd" xmlns:wsp="http://www.w3.org/ns/ws-policy"
    xmlns:wsp1_2="http://schemas.xmlsoap.org/ws/2004/09/policy" xmlns:wsam="http://www.w3.org/2007/05/addressing/metadata" xmlns:wsaw="http://www.w3.org/2006/05/addressing/wsdl"
    xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/" xmlns:tns="http://server/" xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns="http://schemas.xmlsoap.org/wsdl/"
    targetNamespace="http://server/" name="WM_full_serviceService">
  + <wsp:Policy xmlns:wsapw3c="http://www.w3.org/2006/05/addressing/wsdl" wsu:Id="WM_full_servicePortBinding_Wsaw_Addressing_Policy-
      WM_full_servicePortBinding_WSAM_Addressing_Policy-WM_full_servicePortBinding_WSAM_Addressing_Policy">
  - <types>
    - <xsd:schema>
        <xsd:import namespace="http://server/" schemaLocation="http://localhost:8080/WebMiner/WM_full_serviceService?xsd=1" />
      </xsd:schema>
    </types>
  + <message name="getAdvancedResults">
  + <message name="getAdvancedResultsResponse">
  + <message name="getNotAdvancedResults">
  + <message name="getNotAdvancedResultsResponse">
  + <message name="destroySession">
  + <message name="destroySessionResponse">
  + <message name="showTemplates">
  + <message name="showTemplatesResponse">
  - <portType name="WM_full_service">
    + <operation name="getAdvancedResults">
    + <operation name="getNotAdvancedResults">
    + <operation name="destroySession">
    + <operation name="showTemplates">
    </portType>
  + <binding name="WM_full_servicePortBinding" type="tns:WM_full_service">
  - <service name="WM_full_serviceService">
    - <port name="WM_full_servicePort" binding="tns:WM_full_servicePortBinding">
        <soap:address location="http://localhost:8080/WebMiner/WM_full_serviceService" />
      </port>
    </service>
</definitions>
```


Fig. 2. WSDL of the data mining service.

this layer is empty, but this will be developed when we connect our service to any data service in the cloud, for example, Amazon which provides a simple Web Service interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web as well as when we use data mining algorithms which are offered as a Web Service [29,67].

The fifth layer, called presentation layer, is usually not included in discussions about SOA, since SOA decouples the user interface from the components. But, in our opinion it is always present since, in the end, providing an end-to-end solution from an access channel to a service or composition of services is always needed. We explain the implementation developed for our ElWM tool in the following section.

Next, we describe briefly the internal details of the implementation of our service.

4.1. Operating mode

The service works as follow. It offers a set of templates that specify the data which must be sent to the service in order to obtain certain patterns or models that give response to the users' questions. These templates contain the definition of the attributes as well as the mining algorithms which are suitable for obtaining the patterns. Since one of the difficulties which data miners face is the selection of parameters and how these affect the result, the parameters of the algorithms are established by the service itself, by making a previous analysis of the data and/or using other mining algorithms. The definition of these templates is made from a rigorous experimentation in each business domain.

Thus, the user interface, which is built to use this service, must allow the end-user to send the data file or indicate where it is stored and next, invoke the corresponding method getNonAdvancedResults or getAdvancedResults (see Section 4.2) and finally process and show the results obtained. As the service offers the possibility of invoking the template again and changing the parameters, the interface must contemplate this functionality.

4.2. Public functions

As can be observed in Fig. 2 the public functions that the data mining service offers are:

- getNonAdvancedResults: this method generates the data mining model. It has as input parameters: the URL where the data is located, type of data source (database or file), the template to use, a flag indicating if it is the first execution or a second or posterior execution, and another flag which gathers if the end-user requires the pattern with a greater or lesser level of detail. The method returns a URL where the comprised result file is located. This is done to avoid the SOAP messages being very large.
- getAdvancedResults: this method operates in the same way as the previous one although it adds an input parameter which contains the list of the parameters with which the service will execute the mining algorithm. This method is designed to be used by advanced users.
- destroySession: this method, when invoked by the end-user, destroys the session and the resources generated in it.
- showTemplates: this method sends the templates available for each business domain.

4.3. Repository

The repository stores the required information so that the data mining service functions. This is currently implemented in an SQL Server database and includes the following information:

- DM_Algorithm table: it stores the algorithms available in the service.
- XSD_Template table: it gathers the XSD files which define the data file structure to be sent to the service.
- PreprocessTask table: it contains the information about the different pre-processing tasks which can be carried out on the data.
- Template table: it stores the definition of each template. That means the XSD file, the DM algorithms and pre-processing tasks which the service must carry out to generate the model.

```
<!-- Soap Request -->
- <S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/envelope/">
  - <S:Header>
      <To xmlns="http://www.w3.org/2005/08/addressing">http://localhost:8080/WebMiner/WM_full_serviceService</To>
      <Action xmlns="http://www.w3.org/2005/08/addressing">http://server/WM_full_service/getNonAdvancedResultsRequest</Action>
    - <ReplyTo xmlns="http://www.w3.org/2005/08/addressing">
        <Address>http://www.w3.org/2005/08/addressing/anonymous</Address>
      </ReplyTo>
      <MessageID xmlns="http://www.w3.org/2005/08/addressing">uuid:e8cc0933-cd33-433a-96a8-0bfcabf79144</MessageID>
    </S:Header>
  - <S:Body>
    - <ns2:getNonAdvancedResults xmlns:ns2="http://server/">
        <data>http://localhost:1760/data/data.xml</data>
        <source>file</source>
        <template>template_1</template>
        <flag>0</flag>
      </ns2:getNonAdvancedResults>
    </S:Body>
  </S:Envelope>

<!-- Soap Response -->
- <S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/envelope/">
  - <S:Header>
      <To xmlns="http://www.w3.org/2005/08/addressing">http://www.w3.org/2005/08/addressing/anonymous</To>
      <Action xmlns="http://www.w3.org/2005/08/addressing">http://server/WM_full_service/getNonAdvancedResultsResponse</Action>
      <MessageID xmlns="http://www.w3.org/2005/08/addressing">uuid:92dd2375-8a2f-44af-8775-5ce05f64a98c</MessageID>
      <RelatesTo xmlns="http://www.w3.org/2005/08/addressing">uuid:e8cc0933-cd33-433a-96a8-0bfcabf79144</RelatesTo>
    </S:Header>
  - <S:Body>
    - <ns2:getNonAdvancedResultsResponse xmlns:ns2="http://server/">
        <return>"http://localhost:8080/WM/178278789342/compressedResults.zip</return>
        <return xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:nil="true" />
        <return xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:nil="true" />
        <return xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:nil="true" />
        <return xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:nil="true" />
        <return xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:nil="true" />
        <return xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:nil="true" />
      </ns2:getNonAdvancedResultsResponse>
    </S:Body>
  </S:Envelope>
```


Fig. 3. Example of SOAP messages.

As can be deduced, offering new templates only requires adding the corresponding rows in the repository. The WS-DM algorithms must only be recompiled when a new algorithm or pre-processing task is required.

4.4. Security and privacy

Web Services have the characteristic of being stateless. But, many applications such as the typical shopping cart application need to maintain the state or resources of each individual service requester. This is also the case of our service. Therefore, our service creates a session for each end-user and associates it with the files generated as a result (models or patterns) and the parameters of the data mining algorithms used in each request in this session. This allows the service to re-execute the process and refine the pattern or model obtained if the end-user requests it. The session is eliminated when the session is closed or the timeout expires. The use of sessions, in turn, allows the service to maintain privacy, given that an end-user cannot see the data or results generated by another end-user.

Regarding the security, in the near future, the service will offer the possibility of sending and receiving encrypted data files in order to protect sensitive data for which an RSA algorithm proposed by W3C [74] will be used. Additionally, we are also planning to encrypt SOAP messages by means of WS-Security, a flexible and feature-rich extension to SOAP for applying security to Web services.

5. Use of the data mining service: E-learning Web Miner

E-learning Web Miner is an application developed in the University of Cantabria with the aim of helping instructors involved in distance education to discover their students' behavior profiles and models about how they navigate and work in their virtual courses, which are offered in Learning Content Management Systems (LCMSs), such as Blackboard or Moodle.

One of the problems which instructors face is the lack of information about the activity carried out by their students in the course as well as their progress. Although LCMSs provide instructors with certain information, this is limited and not very significant when assessing the teaching-learning process [24]. In general, LCMSs offer a report with summarized access information such as the dates of the first and the last connection, the number of visited pages, the number of read/sent mails and so on for each student; and another report, about the use of resources (announcements, discussions, etc.) with parameters such as number of accesses and time spent on each one. Furthermore, this data is generally shown in a tabular format and not in an intuitive and graphical way so that the instructor,

Fig. 4. ElWM user interface — selection of the mode of operation.

with just a glimpse, can ascertain the students' situation in the course. As a consequence of this, getting a clear vision of the academic progress of each student or group during the course is difficult and time consuming for instructors [17]. Although there are several works in which different alternatives focused on showing graphical reports have been proposed to make this information more understandable such as CourseVis [40], Gismo [43], Moodog [79] and Matep [80], these do not answer questions such as: What are students' profiles according to demographic and navigation information?, how to group students according to their style of learning?, What is the drop-out students' profile?, What are the resources which are frequently used together? or What are the questions in a test which students fail more frequently? In order to answer these questions, the use of data mining techniques is required [56,6]. But as mentioned previously, developing data mining projects is not a trivial task. It requires the skills of being able to map the business goals to the appropriate mining algorithms, choose attributes, perform data transformations, build models and test the results. Furthermore, as Mounir Ben Ayed et al. [3] stated "data mining tools are usually difficult to exploit because most of the end-users are expert neither in computer science nor in statistics". For both reasons, a data mining service for "non-expert data miners" which allows them to discover useful and novel knowledge is necessary.

Next, we show the ElWM application built in the University of Cantabria to answer the instructors' questions and which uses the proposed data mining service.

5.1. ElWM application

ElWM is an application built using standard web technologies and Java programming language.

Its graphical user interface, developed as light client, is implemented with AJAX technology (shorthand for Asynchronous JavaScript and XML) which is a group of interrelated web development techniques used on the client side to create Rich Internet Applications (RIA). As can be observed in Fig. 4, the interface offers two possible forms of use: one for instructors without data mining knowledge in which users only have to send the data file according to the template and request its execution (amateur user) and another in which instructors, before running the process, can additionally establish the parameters of the algorithms (advanced user). Once the mode of operation is selected, the application shows the available templates (see Fig. 5) and, when one is chosen, the application requests the URL where the data file is stored.

Next, ElWM offers the results to the instructor and allows him or her to request a new model by simply indicating if he or she wants a greater or lesser level of detail, this means, if the user wants to observe more or less rules (in the case of using an association rule miner) or a greater or lesser number of clusters (in the case of using clustering algorithms). An example of a page with the result corresponding to the student profile template (see Section 5.2) can be observed in Fig. 6.

5.2. ElWM templates

Currently, ElWM offers three templates: Student profile, Pattern of resources which are frequently used together and Session profile.

Student profile aims to group students according to their activity in the e-learning platform and their demographic data. After an intense and extensive experimentation, we chose as input parameters: gender, age, number of sessions in the course, time spent in the course, average sessions per week, and average time spent per week; and as data mining algorithms for obtaining the patterns: EM (Expectation–Maximization) and SimpleKMeans [74]. EM algorithm is used to determine the number of clusters with which the SimpleKMeans algorithm will be executed (required parameter). We generate the patterns with SimpleKMeans [39] because it is one of the most used in practical problems, its execution is quick and furthermore the results which it offers are easy to understand statistically and graphically. Each cluster is represented by its centroid, which means, the "average" of all its points (average for numerical data and most-frequent value for categorical data). Before the mining process starts, a pre-processing task is carried out in order to evaluate the quality of the data for the process, for example, to eliminate correlated or highly unbalanced data or outliers.

Next, in Fig. 7, we show an example of the result obtained for a data set from a multimedia course taught in the second semester of the 2009/2010 academic year at the largest virtual campus in Spain, called G9 Group, which is composed of 9 Spanish universities, one of which is the University of Cantabria. This course is eminently practical and has the objective of teaching the students how to use a particular multimedia tool called ToolBook. The multimedia course is designed by means of web pages and includes some video tutorials,

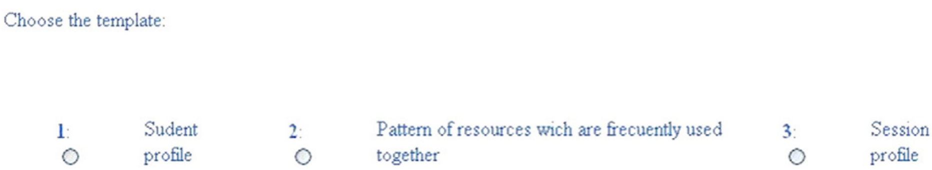

Fig. 5. ElWM user interface — selection of template.

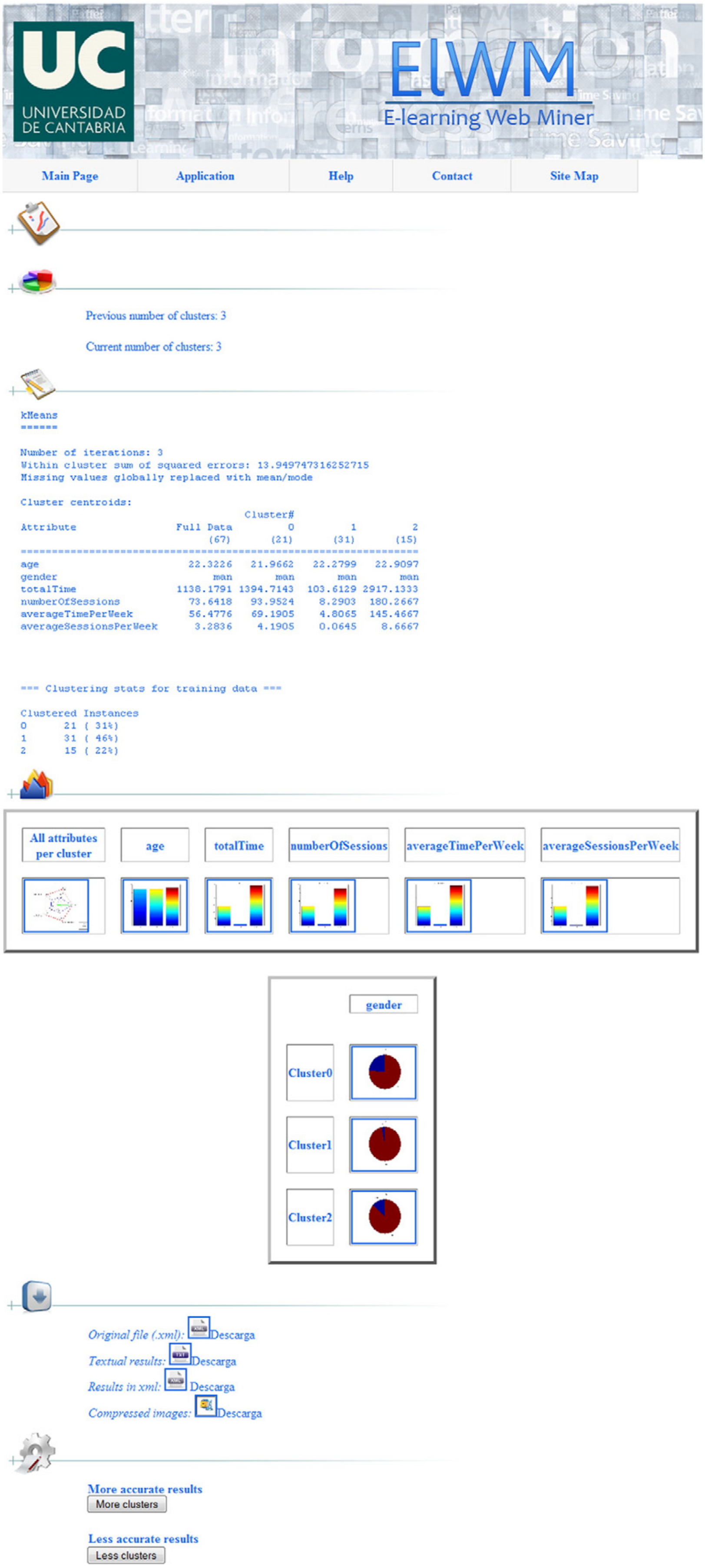

Fig. 6. ElWM user interface — page with the model generated using the student's profile template.

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 13.949747316252715
Missing values globally replaced with mean/mode

Cluster centroids:
                                    Cluster#
Attribute                Full Data         0          1          2
                              (67)      (21)       (31)       (15)
==================================================================
age                        22.3226   21.9662    22.2799    22.9097
gender                         man       man        man        man
totalTime                1138.1791 1394.7143   103.6129  2917.1333
numberOfSessions           73.6418   93.9524     8.2903   180.2667
averageTimePerWeek         56.4776   69.1905     4.8065   145.4667
averageSessionsPerWeek      3.2836    4.1905     0.0645     8.6667



=== Clustering stats for training data ===

Clustered Instances
0      21 ( 31%)
1      31 ( 46%)
2      15 ( 22%)
```

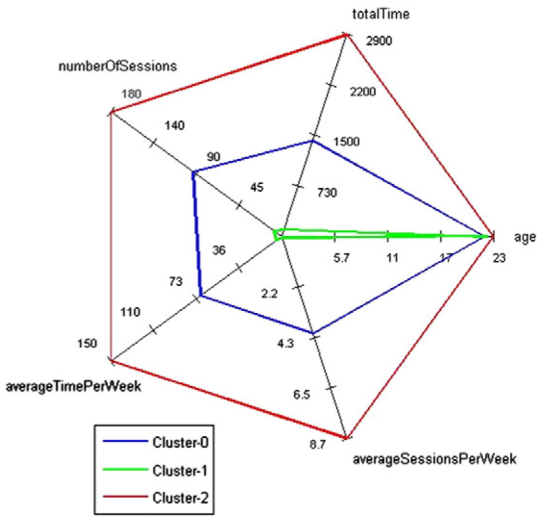

Fig. 7. Student's profile for the multimedia data set.

flash animations and interactive elements. The students must perform 4 exercises, 2 projects and one final exam online. The course is open to all degrees. The number of students enrolled was 80, but only 20% of these delivered all the tasks and 19 students passed the course.

The left-hand side of Fig. 7 shows the textual result obtained from the multimedia data set. As can be observed the service generates three clusters, Cluster-1 with a very low activity which corresponds to students who dropped out in the first days of the course, Cluster-0 which gathers learners with an average activity in time and number of sessions, which is half of the activity carried out by the students collected in Cluster-2. It must be pointed out that there are 13 students who never accessed the course (there are only 67 transactions of 80) and most of the students who enrolled in the course are men.

The same information is drawn graphically on the right-hand side of Fig. 7. This spider graph helps to compare, at a glance, the clusters obtained. It can only represent numerical variables, so that the service offers other graphic results in which the distribution of each attribute in each cluster is shown as can be seen in Fig. 8.

The template about resources which are frequently used together is directed toward the discovery of the resources which are more commonly used together in each session, thus allowing instructors to find out which tools are used more frequently (wiki, chat, forum, etc.) by their students and which ones are basically ignored. This information is very valuable in order to propose tasks according to the learners' learning styles.

The algorithm chosen for this template is the implementation of Apriori developed by Borgelt [4] since it offers a simple rule set

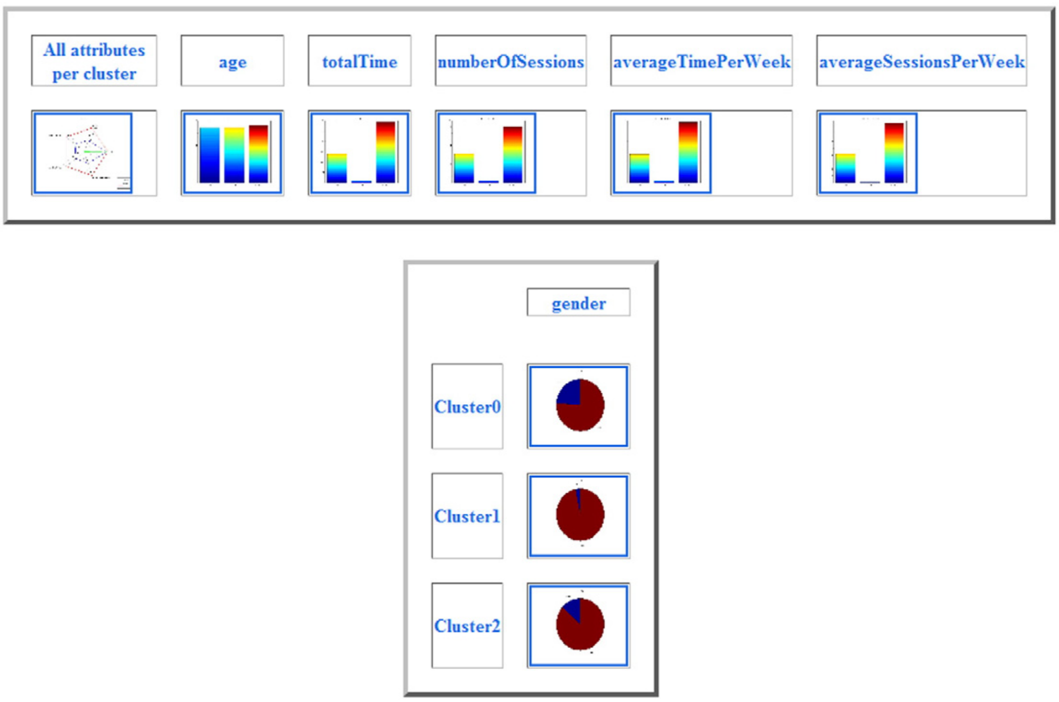

Fig. 8. Graphical representation of the distribution of each attribute in each cluster and the distribution of the value of each attribute in each cluster.

with only one item in the consequent. The algorithm receives a file with the following input parameters: a session identification and a list of the resources used in the session, for example content-pages, mail, forum, chat, and others.

As with the previous example, Fig. 9 depicts the rules obtained by the service using a data set which contains nearly 5000 transactions, one for each learning session.

In this case, the service established the two parameters required by the algorithm, the confidence as 0.7 and the support as 0.10, using a heuristic obtained from the frequent itemset calculation. The instructor can see that the organizer was the tool most often used, followed by the forum (rule_3), assignments (rule_4) and content-pages (rule_2). Furthermore, the instructor can observe that the students visited the forum in the study sessions (rule_10) as well as in sessions of doing tasks (rule_12). So, the instructor can conclude that this resource is suitable for solving problems or doubts since chat and mail were scarcely used. Reading rule_11, the instructor can conclude that students visited the content-pages to do the assignments.

Next, we describe the session profile which helps instructors to better understand the usage pattern of the resources and complements the knowledge provided by the pattern of the resources which are frequently used together, in the sense that it allows the measuring of the level of use of each tool measured in time and hits. The input variables of this template are the number of hits and time spent in each learning session (minutes) in each resource. As there are resources with very low use, the service eliminates those whose use is lower than 1% of the most used resource with the aim of making the pattern more understandable. The algorithm chosen for this template is x-means [51], an extension of k-means which estimates the number of clusters.

Looking at Fig. 10, the instructor can observe that forum and assessment were the more used tools since Cluster-2 and Cluster-3 sum up 89% of the sessions in which practically only these resources and the organizer were used. In both clusters, the students seemed to consult the forum and/or submit an assignment since the time spent and hits done are low. Cluster-1 collects mainly study sessions with an hour of dedication and an average of 27 viewed-pages. Cluster-0 gathers sessions in which learners were developing the tasks and looking up content-pages at the same time. It must be pointed out that in this course the assignments were described in several html pages and that is the reason for having several clicks. This last cluster also gathers the activity in the assessment tool.

In the instructor's opinion, these patterns allow her to gain an insight into the characteristics of her students in relation to the

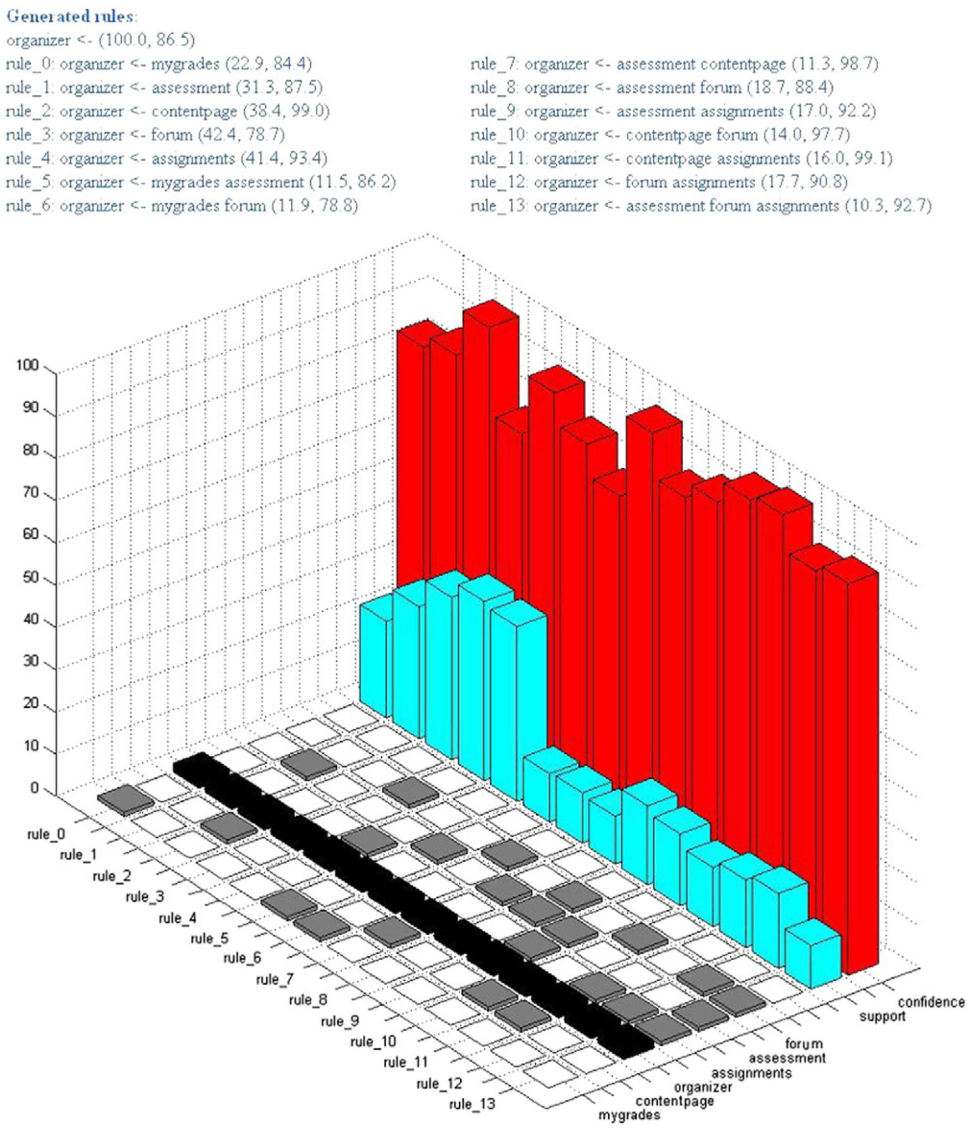

Fig. 9. Resources more frequently used together for the multimedia data set.

Attributes	Cluster 0 (271)	Cluster 1 (256)	Cluster 2 (1051)	Cluster 3 (3356)
time_mail	1.3476	0.4169	0.8051	0.4140
time_forum	4.1622	3.8892	2.0904	0.7084
time_content-page	11.0927	68.6051	1.4502	3.3794
time_organizer	2.6655	6.7933	0.5586	0.7770
time_assignments	20.9470	4.4243	5.2296	0.5603
time_filemanager	0.9801	0.1697	0.1729	0.0128
time_who_is_online	0.1655	0.3321	0.0715	0.0619
time_assessment	4.4437	1.2952	0.6003	0.0566
time_my_grades	0.7185	0.3726	0.4135	0.1552
time_compiler	0.1721	1.9926	0.0347	0.1087
hit_mail	1.0264	0.4760	0.6610	0.3868
hit_forum	10.2682	6.4538	6.5029	1.9123
hit_content-page	3.5033	27.3763	0.6083	1.2411
hit_organizer	5.9304	11.5682	2.0039	1.8634
hit_assignments	5.6490	1.5202	2.4642	0.3511
hit_filemanager	0.5562	0.1217	0.1978	0.0220
hit_who_is_online	0.3973	0.2140	0.1153	0.0837
hit_assessment	2.7152	0.9409	1.1292	0.1642
hit_my_grades	0.6225	0.2546	0.5367	0.2131
hit_compiler	0.0695	0.4280	0.0049	0.0268

Filtered clusters

0	271	(5%)
1	256	(5%)
2	1051	(21%)
3	3356	(68%)

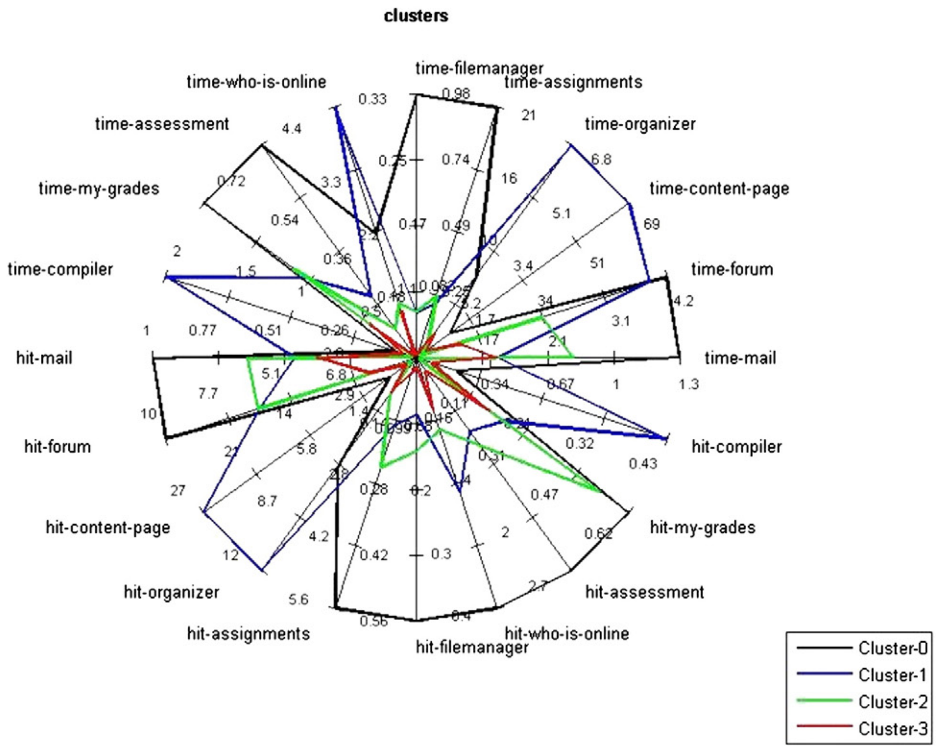

Fig. 10. Session profile for the multimedia data set.

time spent and the use of resources available in the course. Although it is true that the learning process can be carried out without being connected, the interaction of the students with the different resources contains information that can improve their use. This allows her to validate or refute hypothesis used in the design of the learning process.

Currently a new template is being added to the service. This aims to predict the final mark which the student will obtain in the course according to the global activity the student has carried out during the course (total time spent, number of sessions carried out, average number of sessions per week, average time spent per week and average time spent per session). After an intense experimentation, complemented and contrasted with the works published by other authors such as [12,15], we have decided to wrap two Weka classification algorithms, Naïve Bayes and J48. The first one will be used when the sample size is very small (less than 100 instances) and contains numeric attributes, and J48 for data sets with a larger number of instances and/or with the presence of nominal attributes with missing data (very frequent in e-learning data sets). For the selection of these algorithms, we have also taken into account the output which they offer, probability matrices and decision trees respectively, which are easily interpretable and comprehensible to instructors. We will use 10-fold cross validation for estimating generalization performance of the model.

5.3. Other applications

The proposed architecture is easily adaptable to other applications such as survey analysis, customer analysis, market trend, etc. It is only necessary to define the pre-processing tasks and the algorithms to be used and configure the corresponding template and register it in the repository.

This service, at this time, is suitable for small or medium-sized data sets since it is not specifically designed to support Grid requirements and, as is known, the time which data mining algorithms require to process a data set increases exponentially with its size. Thus, we can say that this service is appropriate for small to medium range organizations which are more constrained by the high cost of data mining software and consequently they can use this service without having the costs associated with buying and setting-up software and training their human resources.

6. Conclusions

The delivery of data mining as a service is an emergent necessity, above all for small to medium range organizations which are the most constrained by the high cost of data mining software and the availability of expert data miners to use this software. Until now, the tools deployed as Bi-a-as-Service in the cloud are conceived more for license cost-saving than as a product which can be used directly by end-users without data mining knowledge.

To respond to this necessity, this paper describes the architecture of a data mining service for non-expert data miners which can be delivered as SaaS. Its main characteristic is that it is based on the use of templates that answer certain previously-defined questions. These templates gather the tasks of the KDD process to be carried out on the data set which is sent by the end-user. The templates are defined by the service administrator.

This service is offered as a Web Service which makes it easily accessible from any client application. Furthermore, its extension with other data mining algorithms and visualization tools developed ad-hoc or consumed from a provider in the Internet can be effortlessly incorporated since it is designed following a service-oriented architecture.

This paper also presents ElWM, a web application which uses the data mining service configured for an educational context; in particular, it helps instructors involved in virtual teaching to discover their students' profile and their behavior in the course. The prototype of this web application has been successfully tested in two virtual courses taught in the University of Cantabria. In the instructors' opinions, the tool is very easy to use and the information which it returns is very useful to better understand what is happening in the course and take actions as soon as anomalous behaviors are detected.

Currently, our research is focused on the specification of new templates to be incorporated in the service and consequently, wrapping other data mining algorithms and visualization techniques. The security is another important aspect we are considering. Among other tasks, we have in mind adding encryption techniques in communication messages and data files. Another challenging task is to adapt our architecture to the Open Grid Services Architecture (OGSA) [47] which represents an evolution toward a Grid system architecture based on Web Service concepts and technologies. Lastly, we will study and choose the most suitable cloud environment in which to deploy our solution, for example, Amazon.

Acknowledgment

The authors are deeply grateful to CEFONT, the University of Cantabria department responsible for LCMS maintenance, likewise Elena Álvarez and Rafael Menéndez, the instructors involved in the virtual courses used in our experimentation. The authors would also like to thank the guest editors and anonymous reviewers for their comments that have helped to improve this manuscript. This work has been subsidized by the Spanish Ministry of Science and Technology under TIN2008–05924 project.

References

[1] ADAPA on the Cloud Retrieved April, 2011 from http://www.zementis.com/on-the-cloud.htm.
[2] A. Arsanjani, Service-oriented modeling and architecture. How to identify, specify, and realize services for your SOA Retrieved April, 2011 from http://www.ibm.com/developerworks/library/ws-soa-design1/2004.
[3] M. Ayed, H. Ltifi, C. Kolski, Adel M. Alimi, A user-centered approach for the design and implementation of KDD-based DSS: a case study in the healthcare domain, Decision Support System 50 (2010) 64–78.
[4] C. Borgelt, Efficient implementations of Apriori and Eclat, First Workshop of Frequent Item Set Mining Implementations, Melbourne, 2003.
[5] P. Brezany, I. Janciak, A. Woehrer, A.M. Tjoa, GridMiner: a framework for knowledge discovery on the Grid — from a vision to design and implementation, Cracow Grid Workshop, Cracow, December 12–15, 2004.
[6] F. Castro, A. Vellido, A. Nebot, F. Múgica, Applying data mining techniques to e-Learning problems, in: J. Kacprzyk (Ed.), Studies in Computational Intelligence, Springer-Verlag, 2007, pp. 183–221.
[7] K. Channabasavaiah, K. Holley, E.M. Tuggle, Migrating to a service-oriented architecture, IBM DeveloperWorks Retrieved April, 2011 from https://www.ibm.com/developerworks/library/ws-migratesoa/2004.
[8] M.C. Chen, A.L. Chiu, H.H. Chang, Mining changes in customer behavior in retail marketing, Expert Systems with Applications 28 (2005) 773–781.
[9] Y. Chen, S. Spangler, J. Kreulen, S. Boyer, T. Griffin, A. Alba, A. Behal, B. He, L. Kato, A. Lelescu, C. Kieliszewski, X. Wu, L. Zhang, SIMPLE: a strategic information mining platform for licensing and execution, Proc. of the 2009 IEEE international Conference on Data Mining Workshops, IEEE Computer Society, Washington, DC, 2009, pp. 270–275.
[10] X. Cheng, H. Liu, Personalized services research based on web data mining technology, Second International Symposium on Computational Intelligence and Design, Changsha, 2009.
[11] K. Chine, Scientific computing environments in the age of virtualization, toward a universal platform for the Cloud, Proc. of the 2009 IEEE International Workshop on Open Source Software for Scientific Computation (OSSC), 2009, pp. 44–48.
[12] M. Cocea, S. Weibelzahl, Cross-system validation of engagement prediction from log files, Proc. of the Second European Conference on Technology Enhanced Learning Sustaining TEL: from Innovation to Learning and Practice, 2007, pp. 14–25.
[13] M. Colan, Service-Oriented Architecture expands the vision of Web services Part 2, IBM DeveloperWorks, 2004 Retrieved April, 2011 from http://www.ibm.com/developerworks/webservices/library/ws-soaintro2/.
[14] F. Curbera, Y. Goland, J. Klein, F. Leymann, D. Roller, S. Thatte, S. Weerawarana, Business Process Execution Language for Web Service (BPEL4WS) 1.0 Retrieved April, 2011 from http://www.ibm.com/developerworks/library/ws-bpel August 2002.
[15] G. Dekker, M. Pechenizkiy, J. Vleeshouwers, Predicting students drop out: a case study, Proc. of the 2nd International Conference on Educational Data Mining, 2009, pp. 41–50.
[16] D. Delen, A comparative analysis of machine learning techniques for student retention management, Decision Support Systems 49 (2010) 498–506.
[17] I. Douglas, Measuring participation in internet supported courses, in Proc. of the 2008 International Conference on Computer Science and Software Engineering, IEEE Computer Society, Washington, DC, 2008, pp. 714–717
[18] J. Erickson, K. Siau, Web services, service-oriented computing and service-oriented architecture: separating hype from reality, Journal of Database Management 19 (2008) 42–54.
[19] T. Erl, Service-Oriented Architecture: Concepts, Technology, and Design, 1st ed. Prentice Hall, 2005.
[20] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining Boston, AAAI/MIT Press, 1996.
[21] L. Feng, J. Yu, H. Lu, J. Han, A template model for multidimensional inter-transactional association rules, The VLDB Journal 11 (2002) 153–175.
[22] Good data Retrieved April, 2011 from http://www.gooddata.com/.
[23] D. Guedes, W. Meira Jr., R. Ferreira, Anteater: a service-oriented architecture for high-performance data mining, IEEE Internet Computing 10 (2006) 36–43.
[24] A. Haira, D. Birant, A. Kut, Improving quality assurance in education with web-based services by data mining and mobile technologies, Proc. of the 2008 Euro American Conference on Telematics and information Systems, ACM, New York, NY, 2009, pp. 1–7.
[25] R. Hijon, A. Velázquez, E-learning platforms analysis and development of students tracking functionality, in: E. Pearson, P. Bohman (Eds.), Proc. of World Conference on Educational Multimedia, Hypermedia and Telecommunications, Chesapeake, 2006, pp. 2823–2828.
[26] S.M.S. Hosseini, A. Maleki, M.R. Gholamian, Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty, Expert Systems with Applications 37 (2010) 5259–5264.
[27] X. Hu, A data mining approach for retailing bank customer attrition analysis, Applied Intelligence 22 (2005) 47–60.
[28] H. Jiawei, K. Micheline, Data mining: concepts and techniques, Elsevier Inc., San Francisco, 2006.

[29] L. Joita, O.F. Rana, F. Freitag, I. Chao, P. Chacin, L. Navarro, O. Ardaiz, A catallactic market for data mining services, Future Generation Computer Systems 23 (2007) 146–153.
[30] R. Kerber, H. Beck, T. Anand, B. Smart, Active templates: comprehensive support for the knowledge discovery process, Proc. KDD-98, California, 1998 pp. 244–248.
[31] J.U. Kietz, F. Serban, A. Bernstein, S. Fischer, Data mining workflow templates for intelligent discovery assistance and auto-experimentation, in Proc. ECML/PKDD10 Workshop on Third Generation Data Mining: Towards Service-oriented Knowledge Discovery, Catalonia, Spain, 2010.
[32] S.Y. Kim, T.S. Jung, E.H. Suh, H.S. Hwang, Customer segmentation and strategy development based on customer lifetime value: a case study, Expert Systems with Applications 31 (2006) 101–107.
[33] R. Kimball, M. Ross, The Data Warehouse Toolkit, 2nd ed. John Wiley & Sons, New York, 2002.
[34] Y.S. Lee, S.J. Yen, Incremental and interactive mining of web traversal patterns, Information Sciences 178 (2008) 287–306.
[35] A.C.J. de Leeuw, Oranisaties: management, analyse, ontwerp en verandering, Van Gorcum, Assen, 1982.
[36] Lityxiq Retrieved April, 2011 from http://www.lityxiq.com.
[37] J. Liu, Smart shopper: an agent-based web-mining approach to Internet shopping, IEEE Transactions on Fuzzy Systems 11 (2003) 226–237.
[38] Logixml Retrieved April, 2011 from http://www.logixml.com/.
[39] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability, Berkeley, 1967, pp. 281–297.
[40] R. Mazza, V. Dimitrova, CourseVis: a graphical student monitoring tool for supporting instructors in web-based distance courses, International Journal of Human-Computer Studies 65 (2007) 125–139.
[41] P. Mell, T. Grance, The NIST Definition of Cloud Computing Version 15, 2009 Retrieved April, 2011 from http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc.
[42] Microstrategy Retrieved April, 2011 from, http://www.microstrategy.com/.
[43] C. Milani, R. Mazza, GISMO: a graphical interactive student monitoring system for Moodle Retrieved April, 2011 from, http://gismo.sourceforge.net2007.
[44] E. Newcomer, G. Lomow, Understanding SOA with Web Services, Addison Wesley, 2005.
[45] R. Nisbet, J. Elder IV, G. Miner, Handbook of Statistical Analysis and Data Mining Applications, Academic Press, 2009.
[46] E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen, Xin Sun, The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature, Decision Support System 50 (2011) 559–569.
[47] Open Grid Service Architecture Retrieved April, 2011 from, http://www.globus.org/ogsa/.
[48] D.L. Olson, D. Delen, Y. Meng, Comparative analysis of data mining methods for bankruptcy prediction, Decision Support Systems 52 (2012) 464–473.
[49] Panorama Retrieved April, 2011 from, http://www.panorama.com/powerapps/.
[50] M.P. Papazoglou, W. van den Heuvel, Service-oriented architectures: approaches, technologies and research issue, The VLDB Journal 16 (2007) 389–415.
[51] D. Pelleg, A. Moore, A. Xmeans, Extending SimpleKmeans with efficient estimation of the number of clusters, Proc. of the Seventeenth International Conference on Machine Learning, Morgan Kauffman, San Francisco, 2000, pp. 727–734.
[52] G. Piatetsky-Shapiro, Kdnuggets web site Retrieved April, 2011 from, http://www.kdnuggets.com/index.html.
[53] Pivotlink Retrieved April, 2011 from, http://www.pivotlink.com/.
[54] V. Podpecan, M. Jursic, M. Zakova, N. Lavrac, Towards a service-oriented knowledge discovery platform, Workshop on Explorative Analytics of Information Networks at ECML PKDD, Bled, Slovenia, 2009.
[55] Rightscale-BI Retrieved April, 2011 from, http://www.rightscale.com/lp/bi-stack.php.
[56] C. Romero, S. Ventura, Educational data mining: a survey from 1995 to 2005, Expert Systems with Applications 33 (2007) 135–146.
[57] D. Romero Morales, J. Wang, Forecasting cancellation rates for services booking revenue management using data mining, European Journal of Operational Research 202 (2010) 554–562.
[58] J. Rushing, R. Ramachandran, U. Nair, S. Graves, R. Welch, H. Lin, ADaM: a data mining toolkit for scientists and engineers, Computers & Geosciences 31 (2005) 607–618.
[59] SAP On-Demand Solutions Retrieved April, 2011 from, http://www.ondemand.com/.
[60] A. Shaikh Ali, Federated Analysis Environment for Heterogeneous Intelligent Mining Project Retrieved April, 2011 from, http://users.cs.cf.ac.uk/Ali.Shaikhali/faehim/index.htm.
[61] M.J. Shaw, C. Subramaniam, G.W. Tan, M.E. Welge, Knowledge management and data mining for marketing, Decision Support Systems 31 (2001) 127–137.
[62] Smart Analytics System Retrieved April, 2011 from, http://www.ibm.com/ibm/cloud/cloudburst/.
[63] K.A. Smith, R.J. Willis, M. Brooks, An analysis of customer retention and insurance claim patterns using data mining: a case study, The Journal of the Operational Research Society 51 (2000) 532–541.
[64] SOA Manifesto Working Group, The SOA Manifesto Retrieved April, 2011 from, www.soa-manifesto.org.
[65] Software & Information Industry Association, Software-as-a-Service: A Comprehensive Look at the Total Cost of Ownership of Software Applications Retrieved April, 2011 from, http://www.winnou.com/saas.pdf 2006.
[66] V. Srinivasa Rao, N. K. Nagewara Rao, E. K. Kumari, Cloud Computing: An Overview. Journal of Theoretical and Applied Information Technology 10. Retrieved April, 2011 from, http://www.jatit.org/volumes/research-papers/Vol9No1/10Vol9No1.pdf
[67] D. Talia, P. Trunfio, O. Verta, The Weka4WS framework for distributed data mining in service-oriented Grids, Concurrency and Computation: Practice and Experience 20 (2008) 1933–1951.
[68] I.-H. Ting, H.-J. Wu, Web mining applications in e-Commerce and e-Services, Series: Studies in Computational Intelligence, Springer, 2009.
[69] C. Tsai, M. Tsai, A dynamic web service based data mining process system, Fifth International Conference on Computer and Information Technology (CIT'05), 2005, pp. 1033–1039.
[70] J. Vanschoren, H. Blockeel, Stand on the shoulders of giants: towards a portal for collaborative experimentation in data mining, Proc. of Service-oriented Knowledge Discovery in ECML/PKDD, Bled, 2009, pp. 88–99.
[71] J. Vanschoren, B. Pfahringer, G. Holmes, Learning from the past with experiment databases, Working Paper Series, 08/2008.
[72] D. Veste, EXCERPT Worldwide Business Intelligence Tools 2009 Vendor Shares, IDC, 223725E1, 2010.
[73] W3C Recommendation: XML Encryption Syntax and Processing — RSA Retrieved April, 2011 from, http://www.w3.org/TR/2002/REC-xmlenc-core-20021210/Overview.html#rsa-1_5.
[74] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, second ed. Morgan Kaufmann, 2005.
[75] P.C. Wong, P. Whitney, J. Thomas, Visualizing association rules for text mining, Proc. of IEEE Symposium on Information Visualization, IEEE Computer Society, San Francisco, 1999, pp. 120–128.
[76] Yahoo Search Web Services Retrieved April, 2011 from, http://developer.yahoo.com/search/.
[77] Q. Yang, X. Wu, 10 challenging problems in data mining research, International Journal of Information Technology & Decision Making 5 (2006) 597–604.
[78] J. Yeh, T. Wu, C. Tsao, Using data mining techniques to predict hospitalization of hemodialysis patients, Decision Support Systems 50 (2011) 439–448.
[79] H. Zhang, K. Almeroth, A. Knight, M. Bulger, R. Mayer, Moodog: tracking students' online learning activities, World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED MEDIA), Vancouver, Canada, 2007.
[80] M.E. Zorrilla, E. Álvarez, MATEP: monitoring and analysis tool for e-Learning platforms, Proc. 8th IEEE International Conference on Advanced Learning Technologies, Santander, 2008.
[81] M.E. Zorrilla, D. Marín, E. Álvarez, Towards virtual course evaluation using web intelligence, LNCS 4739, Springer-Verlag, Berlin Heidelberg, Las Palmas, 2007, pp. 392–399.

Marta Elena Zorrilla Pantaleón is an Assistant Professor in Computer Science at the University of Cantabria (Spain). She earned her bachelor degree in Telecommunication Engineering and PhD in Computer Science at the University of Cantabria in 1994 and 2001 respectively. She has participated in and managed more than 20 research projects, most of them with companies, and she is an author of a database book and more than 40 works published in international journals, books and conferences. Her research interests are the design and development of information systems and intelligent systems for companies; and, inside the educational area, the application of data mining techniques and OLAP technologies in order to analyze and improve web-based learning sites.

Diego García-Saiz is a student in Computer Science at the University of Cantabria (Spain) since 2005. Currently he is working in educational data mining field as a research scholar at Mathematics, Statistics and Computing department. He has also participated in Real-Time Systems projects at Real-Time Computing department at University of Cantabria and he is an author of 2 works presented in international conferences. His research interests are software engineering, databases and data mining.