A Data Mining Engine based on Internet

J. Sirgo[1], A. López[1], R. Jáñez[1], R. Blanco[1], N. Abajo[2], M. Tarrío[2], R. Pérez[2]
[1] University of Oviedo. Department of Electrical Engineering. C. Viesques, Gijón, 33204, Asturias, Spain
Tlph.: +34 985 18 20 83. Fax: +34 985 18 20 68
[sirgo,antonio,rjanez,raguirre]@isa.uniovi.es
[2] Aceralia, Arcelor Group. Research and Development Center. Apdo. 90, Avilés, 33480, Asturias, Spain
Tlph.: +34 985 12 64 04. Fax: +34 985 12 63 75
[nabajom,mtarriom,rperezc]@aceralia.es

Abstract

Data mining has emerged over the last years as a way to discover valuable information from huge amounts of data. In the industrial environment this data is usually stored at several local databases spread over the factory installations. Under this context of application, Data Mining studies need not only a set of algorithms to analyze the data, but also a way to administer the data under study. Up to the moment, available commercial solutions for data mining projects are rather closed in nature. This paper describes a prototype for a new data-mining architecture based on Internet technology developed around an application server.

Index Terms

Data Warehousing, Data Mining, Open Architectures, Internet-based Data Processing

I. Introduction

Nowadays, computer-based process monitoring is standard practice in modern industry. As a parallel effect, large quantities of data related to the process are gathered and stored, usually employing a relational database as the logical support. In a relatively short time, the volume of data becomes difficult to handle by usual methods. But the database may contain important information about the process history: behavioral patterns that could be of great interest in order to optimize the process, understand it, or simply analyze its behavior over time.

To overcome this problem, numerous strategies have been developed over the years to help in the extraction of knowledge from huge amounts of data. These strategies are known as KDD ("Knowledge Discovery in Databases"), a term that defines the global process of extracting knowledge from low level data. As a key part of these strategies, data mining [2], [3] is defined as the extraction of patterns from observed data, usually involving steps such as those described as the central part of the CRISP-DM Methodology [1]:

- Data collection and preparation.
- Data mining algorithm selection and model building.
- Evaluation of the results.

There are many commercial tools that cover these steps at differing degrees of detail, but in general these kits are closed solutions (they cannot be altered), making it difficult to extend their capabilities with new algorithms and methods.

Another problem that arises with monolithic applications is the client software distribution. In environments with non homogenous computers (different hardware, operating systems, etc.) and non homogeneous users, it is very difficult or almost impossible to design a plan to keep client software updated. Once again, commercial solutions do not facilitate this task. They are based in general on monolithic applications that need to be installed locally at any computer where they will be run.

Under this state of the art, the University of Oviedo and Aceralia have been faced with the development of a data mining service for the company during the last three years. The work was supported under the CECA project QDB. Once the structure of the company was analyzed and the above problems presented, the decision was taken to develop a data mining web-based application server. This paper briefly describes such product, focusing on the main difficulties found during its construction and the solutions adopted.

II. Structure of the Paper

After the previous introduction, the paper continues (Sect. III) with a more detailed analysis regarding the motivation to develop a new architecture for data mining, in place of using a commercial tool. Next, in Sect. IV the architecture of the prototype is shown, analyzing different hardware and software constructive details. The paper continues discussing the main functions of the system, dividing them between the different kinds of users that access to it (Sect. V). At last, the paper finishes (Sect. VI) with a discussion about the strong and weak points of the new approach and describing future trends of work.

III. Motivation of the Work

Aceralia is a steel manufacturing company included in the Arcelor Group. Installations (called plants), are spread over a wide geographical region. At different plants, process data is stored in local databases, mainly for process monitoring and control.

Today's competitive market necessitates the use of a great deal of effort in the enhancement of production processes, from the point of view of maximizing production while enhancing the final quality. In the beginning, data mining techniques were applied in Aceralia on local databases. However, better results could be obtained if global information were used. Therefore, local databases need to be merged and linked, thus providing information about the product during each of its transformation stages (from the input material to the final product).

As a solution, a centralized point was devised (something like a data mart, a process-oriented data warehouse) where such information could be available for users. The goal was to obtain an environment that would allow the smooth integration of a huge quantity of data from local databases and for different types of users, from very specialized researchers to simply those who want to see graphs on data relationships. Also, an environment had to be provided for access to and exploitation of the data mart.

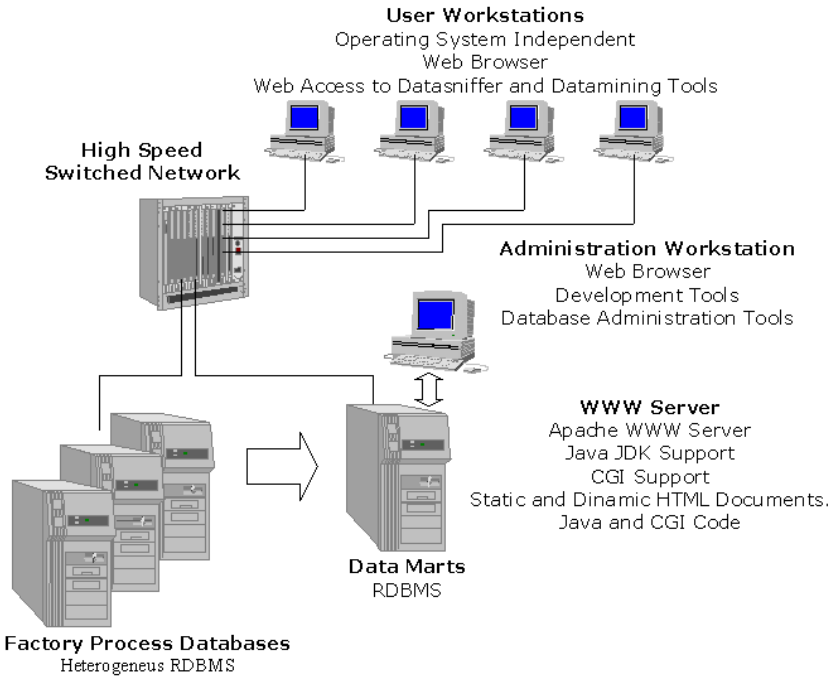


Fig. 1. Developed System Architecture. Workstation access to the kernel of the system is provided through a Web server using an independent navigator. The server provides access to different functions of the system acting on data gathered from the different factory process databases.

IV. Architecture of the Proposed System

The developed system, called **QDB** (see Fig. 1), is a client/server platform where the client is a java-enabled Web navigator. A Web server receives requests from the users serving them as necessary. This way, no specific client software is needed and users always have access to the latest software release.

QDB functions reside mainly in the server by means of servlets or CGI components. Sometimes, for minor functions, Applets are used, but they are mainly avoided in order to discard computations from the workstations since their potential is unknown at the beginning. For every operation, the user must fill in a form for the client with the necessary information. This information is processed in the server and finally the resultant information is loaded into the navigator. Every task is performed using a uniform interface in order to diminish the application management learning time.

Data from local databases is loaded into a relational database engine installed on the server. To provide homogeneous access to such plant databases from QDB, a link must exist to allow access from the server. Of course, access to such links and to external databases is consequently secured by means of normal database security policies.

Data mining algorithms implemented are from heterogeneous sources. Some of them were specifically developed for the project, others are commercial java-implemented products and others are freeware components also developed in Java (sometimes modified in order to fit the architecture).

V. System Functionalities

The main functions agree on the principal user groups that access the system: database administrators for administrative tasks and researchers for data exploiting tasks. Of course, access is provided to general users in order to visualize data from the data mart and print reports.

Every task is carried out through a web form specifically designed for that purpose. Departing from a common skeleton (see Fig. 2), controls are added on the work area to set up the necessary parameters for each specific task and to display the results.

A. Administrator Capabilities

Administrators are expert users who has direct access through the QDB interface to the plant databases, and administers data marts and users. Their main action is the creation of data marts as a result of a meshing of the plant database information. This operation is made up of three steps using a component called *DataSniffer*:

- Select tables and fields to load from the plant databases.
- Relate selected tables in order to create a consistent data mart (set up primary and foreign keys for all the selected tables).
- Launch the data load.

B. Data Miner Capabilities

Data miners are those who carry out studies. They apply data mining techniques to information loaded into the data mart to solve different problems. CRISP-DM Methodology [1] was taken as the reference model.

A common analysis is usually divided into a set of steps:

- Select the variables for the study contained in different columns of the data mart tables.
- Clean and filter the data.
- Apply data mining algorithms.

At the moment, only a few data mining elemental algorithms have been implemented. Fig. 3 shows the interface for one of them: a clustering analysis.

VI. Conclusions and Future Work

The system is still in a test phase. Nevertheless, preliminar tests allow us to settle some conclusions:

- Data mart generation from local databases overloads the communication network. Nevertheless, this is a weak point not only for the system developed, but for every data

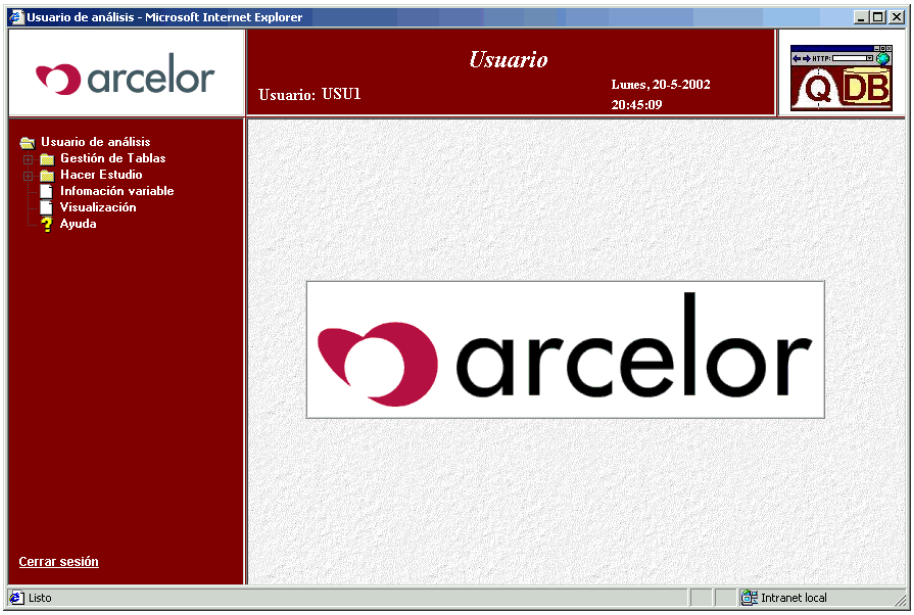


Fig. 2. User Interface. Left part: menu option tree (different for each kind of user). Right part: work area. A specific form is loaded depending on the menu option selected.

mart/data warehouse product. In order to reduce the download time as much as possible, queries to plant databases have been optimized.
Regarding the rest of operations, they do not suposse usually a high load of the network. Nearly all are based on a few parameters that are sent to the server and have as a result a new reduced set of parameters (remember that algorithms run on the server, where data is stored locally). Maybe the exception are graphical representations, where the traffic load increases substantially.

- Data management through Java-based tools is not the fastest approach, but it is a common approach useful for the selected platform of implementation.
- All the computational effort is discharged onto the master. For this, a powerful server is necessary.
- Software maintenance, distribution and scalability are assured, being new capabilities added to the system immediately available for users of the system.
- The system allows easy expansion and enhancing capabilities by means of new implemented algorithms. Moreover, products from third companies, depending of their possibilities, could be added to the system through software gateways.
- All the users interact with the system through a homogeneous user interface. This way the learning curve is smoothed.

As future work, we are planning, first at all, to perform intensive performance analysis, discovering this way new weak points and proposing solutions to them. This is not an easy task because many users need to be involved.

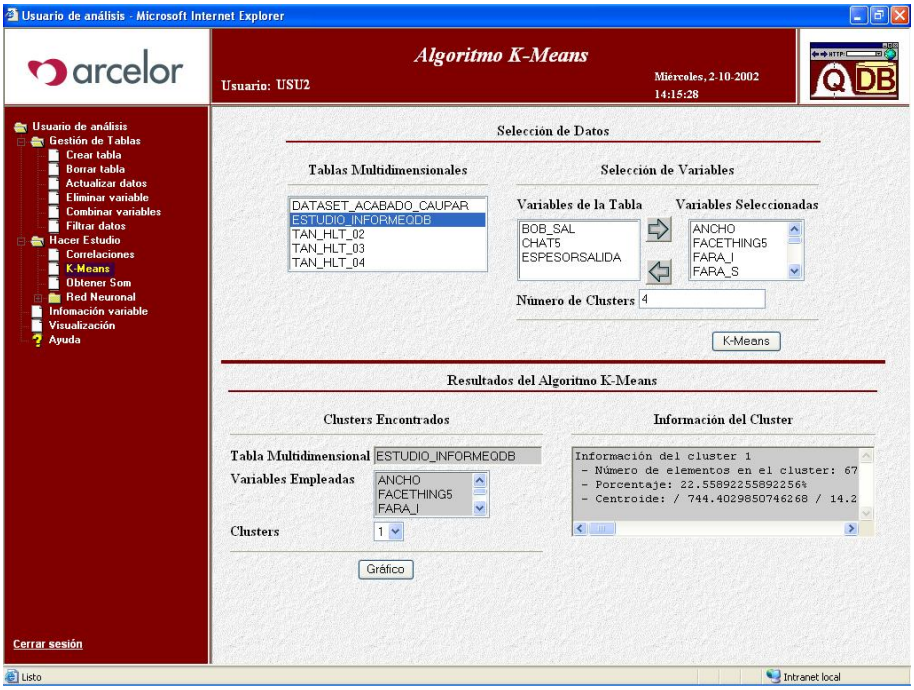
One of the weakest points will be the communications network overload. Up to the moment, we have we have concentrated at building a common framework for data mining, useful to be easy expanded and maintained, not paying much attention at this point. In order to get a useful product, this point must be studied in deep and different alternatives must be tested, such as caching policies.

Another problem is the master work load. Regarding this point, we are planning to set up different masters in a kernel to perform operations in a transparent way for the user.

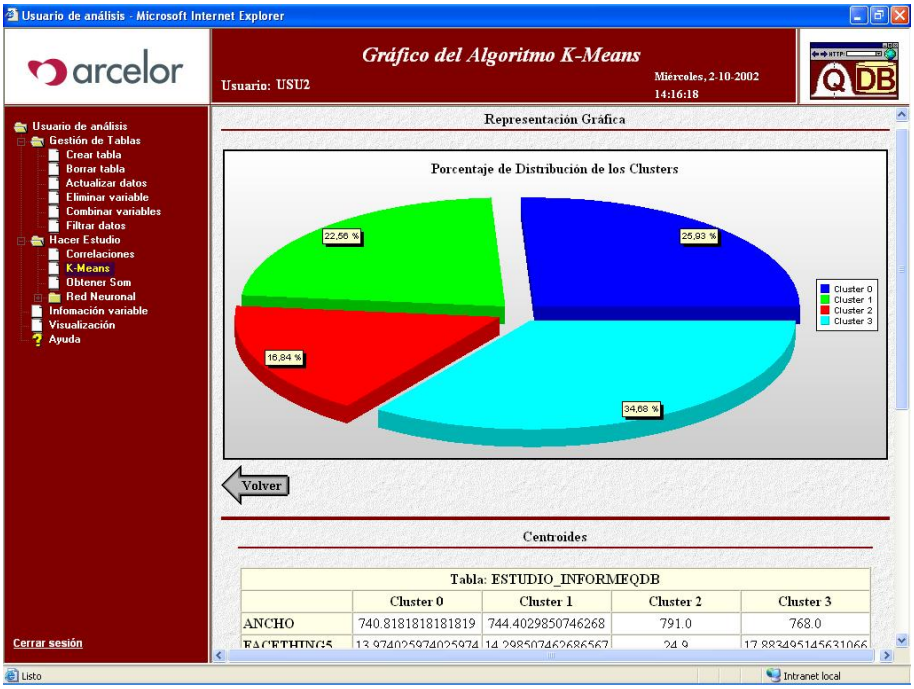
The system can also be expanded with new data analysis algorithms, constructing this way a more complete data mining engine.

REFERENCES

[1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *Crisp-dm 1.0. step-by-step data mining guide*, http://www.crisp-dm.org.
[2] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in knowledge discovery and data mining*, The M.I.T. Press, 55 Hyward Street, Cambridge, MA, 02142, 2000.
[3] J. Han and M. Kambere, *Data mining: Concepts and techniques*, Morgan Kaufmann, 340 Pine St, San Fransisco, CA 94104, USA, 2000.



(a) K-mean Clustering Form. Cluster variables are selected and the desired number of clusters is specified (upper part). In the lower part, clusters found are shown to the analyst.



(b) Clustering results can also be analyzed through a graphical representation.

Fig. 3. User Interface for Data Mining K-mean Analysis.