A Web Mining System

JOSE AGUILAR
CEMISID, Departamento de Computación
Facultad de Ingeniería, Universidad de Los Andes
La Hechicera, Mérida
VENEZUELA

Abstract: - The Web Mining arises like an appropriate tool to exploit the derived knowledge of the web-user interaction, describing models that use patterns and characterize the profiles of the different groups of users which use Internet. To achieve this, currently there are numerous techniques. Some of these techniques are integrated in this work to build a Hybrid System of Web Mining that allows extracting useful information of the web users. Specifically, three techniques of the area of Web Mining were used: Sequential Patterns, Path Analysis and Cubes. The System obtains a group of access patterns from the users to a website, to arrange them in a multidimensional structure, called Cube. Using that, the system can discover correlations between the web pages and users' groups, behaviors of the web users, among other things.

Key-Words: - Web Mining, Sequential Patterns, Path Analysis, Cubes, Pattern Recognition, Data Mining

1 Introduction

The explosive growth of Internet has made more necessary for the users to use automatic tools to find, to extract, to filter and to evaluate the available resources of information over it. There are powerful search tools to find information for category or for content, such as Yahoo, Google, etc. For these searches we need to introduce keywords, and they determine the web pages that contain these words, trying to satisfy the user's requirements. Many times, these consultations bring inconsistence, or documents that fulfill the search approach but not the users interest [2, 6, 10, 11, 14, 15, 16].

There is necessity of having new technologies that help us in our search processes and, even more, of technologies that help us to use the content of the web more efficiently [5, 6, 8]. For this reason, in the last years a series of techniques that allow the advanced processing of data on the Internet have been developed. These techniques carry out a depth analysis in an automatic way, and they belong to an area denominated Web Mining [1, 3, 6].

The web mining is an area that involves the use of techniques based on the data mining, guided for the discovery and automatic extraction of information, of documents and services in the web [4, 5, 6, 11, 12]. The web mining provides tools so that the user can discover and exploit the implicit knowledge in the web.

The main axis of the approach proposed in this work involves the use of techniques of the area of Web Mining (Cubes, Sequential Patterns, Path Analysis), in order to propose a System of Web Mining that is a hybrid of them, in such a way of exploiting the advantages of each one. Our Hybrid System of Web Mining can be used to extract the useful information for the web users, such as: correlate between the web pages and users' groups, behaviour of the users when navigating for Internet, cluster of pages according to the users, among others.

This paper is organized as follows, in section 2 we present the web mining techniques that we are going to use in our proposition. Then, we present the design of our system. Section 4 shows a case of study, and finally, the conclusions are presented.

2 Theoretical Aspects

2.1 Web Mining

The Web Mining can be defined as the automatic analysis and discovery of useful information from documents and services of the web [3, 5, 6]. The Web Mining is an area that involves the use of techniques based on Data Mining, but now guided to the discovery and analysis of the information in the web.

The data mining is a generic term that includes the techniques and tools used to extract useful information from big databases. It arises as an important concept in the data analysis, mainly in

environments where there are multitudes of data and it is necessary to extract all the useful information [12].

The techniques of Web Mining can be used to access more efficient to the information contained in the web, of direct or indirect form. Normally, we can group the web mining approaches in three types of mining [3, 6, 7]:

- *Web Content Mining*: it refers to the automatic search of information and extraction of knowledge, based on the content and the descriptions of the documents in the web.
- *Web Structure Mining*: it refers to the process of inferring knowledge based on the organization and the references or connections among documents of the web.
- *Web Usage Mining*: it is a type of Web Mining that refers to the discovery and analysis of the access patterns or the users' habits, which are extracted from the implicit information of their activities.

2.2 Web Mining Techniques

There are a lot of techniques of web mining [1, 2, 3, 6, 7, 11]. In this work we are going to present only the techniques that we are going to use in our system:

- *Cluster and Classification*: The cluster techniques identify and distribute similar individuals' behaviours in homogeneous groups. Once discoveries the profiles of each group, the characteristics of each one of them can be used to carry out a classification.
- *Rules of Association*: The association rules can be seen as the identification of actions or facts that, being initially independent, they happen in a combined or associate way. The considered facts can be characteristics or behaviours observed in the individuals.
- *Path Analysis*: This technique supposes the generation of directed graphs, which represent the relationships among the web pages. The web pages are the nodes of the graphs, and the connections among the pages are the directed arcs among nodes. They can also be defined like graphs with arcs which represent the similarity among pages, or arcs that show the number of users that go from a page to another.
- *Sequential patterns*: It is a historical of transactions in a web server, where the visit of a client is stored for a period of time. The problem of discovering sequential patterns of access is based on identifying the group of more frequent accesses in a group of transactions or visits in a giving time.
- *Cubes*: a cube of data is a type of multidimensional array that allows the users to explore and to analyze a collection of data from different perspectives. From a structural perspective, the cubes of data are composed of two elements: dimensions and measures. The dimensions are categories that describe the studied factors for their analysis, and the measures are the values of the data stored in that structure.

3 Design of the System

The system of web mining is divided in three subsystems. Each subsystem describes one of the processes that are made with our system: extracted the information from the web, to arrange the information from the web to web mining tasks, and to answer to the consultations of the users of our system. At the following, we are going to describe the different modules of everyone (see figure 1).

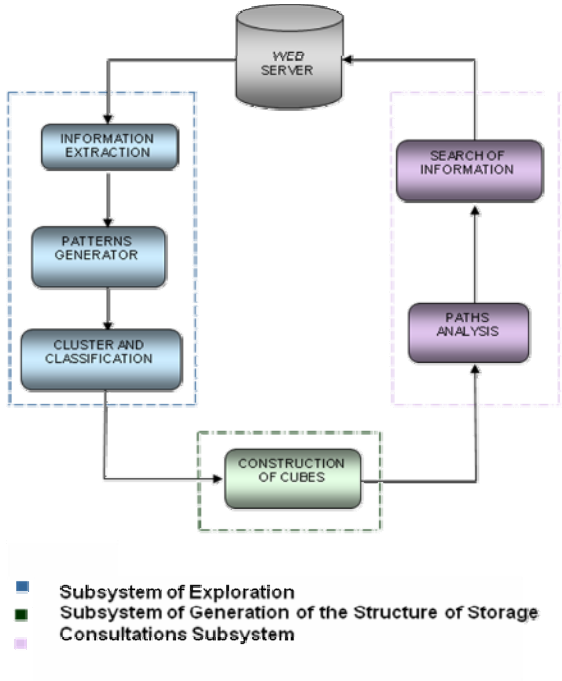
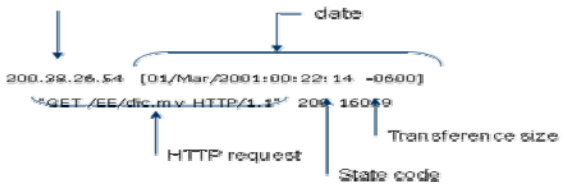


Fig. 1. Our Web Mining System

3.1 Subsystem of Exploration

This subsystem basically makes the extraction of information stored in the web server. Later, it carries out tasks of cleaning on these gathered data. Once obtained these data, it carries out the generation of access patterns of the users', and finally, it classifies all the found patterns. It is composed of three modules:

INFORMATION EXTRACTION MODULE: The log of accesses file is the main source of information on the visitors. In this file is registered all and each one of the transactions between the visitor's navigator and the web server. For example, a line of the Apache log file (http://www.apache.org/) is the following:



Once gathered these files (logs files), our system carried out tasks of cleaning. This way, it eliminates all the unnecessary information to build the access sequence from the user to the website. The sequences or paths consist of each one of the pages visited by an user in the order in that it visited them in a given moment.

PATTERNS GENERATOR MODULE: Starting from the information gathered in the previous module, it applies the algorithm that generates the sequential patterns of the users. This is an algorithm that allows detecting the groups of websites more frequently visited for the users in a given moment. The algorithm has two steps:

- The first step consists on calculating all the access sequences of different lengths to each site,
- The second step selects those groups whose access frequency overcomes a minimum value (threshold) defined by the user. For example: be a group of users U = {U1, U2, ..., U9}, where each Ui is constituted by a group of paths (access sequences), where Li represents the pages visited by each user in the website, and the frequency threshold is 3.

U1: L1, L2, L5	U2: L2, L4	U3: L1, L3
U4: L1, L2, L4	U5: L1, L3	U6: L2, L3
U7: L1, L3	U8: L1, L2, L3, L5	U9: L1, L2, L3

In the first step of the algorithm each path is member of the candidates group and the occurrences number is calculated for each one in all the users (see table 1). Next, it proceeds to eliminate those paths that don't fulfill the frequency threshold. Finally, the group of more frequent sites is obtained.

CLUSTER AND CLASSIFICATION MODULE: To make the cluster of the found patterns, it is necessary to establish an approach to determine the similarity that exists among them. In this case, the relationship of similarity is given by the keyword that defines the content of each one of the pages visited by the user in a path. We could use others similarity relationships defined in the literature [9, 11, 13, 14, 15]. To achieve this, an algorithm is executed that find the four words more frequent in each page. For example, for the pattern: /cursos/index.html /cursos/pgcomp/pgcomp.html, the keywords of each site is:

/cursos/index.html ==> Faculty 10, Engineering 15. University 22, Andes 22.
/cursos/pgcomp/pgcomp.html ==> Computing 15, Data 14Advanced 13, Engineering 12.

Then, the occurrence of each keyword is counted in the path. For example:

Faculty = 1 **Engineering = 2** University = 1
Andes = 1 Computing = 1 Data = 1 Advanced = 1

Finally, it is selected as keyword for this pattern "Engineering", because it is the only one that appears in both sites.

Table 1. Candidate paths and theirs access frequencies

Paths	Frequency
l_1	**7**
l_1 l_2	**4**
l_1 l_2 l_5	1
l_2 l_5	1
l_2	2
l_2 l_4	2
l_1 l_3	**3**
l_1 l_2 l_4	1
l_2 l_3	**3**
l_1 l_2 l_3	2
l_1 l_2 l_3 l_5	1
l_2 l_3 l_5	1
l_3 l_5	1

3.2 Subsystem of Generation of the Structure of Storage

The cubes are built with each group of sequential patterns. Each cube stores all the sequential patterns that correspond to one keyword. To define their structure, we need to define the dimensions and measures that integrate each cube:

Dimensions:
Frequency: In this axis the data are stored according to the access frequency of each sequential pattern.
Length of the pattern: In this dimension is takes into account the pattern's length, that is, the number of sites that constitute the pattern.
Sequential pattern: In this dimension the sequential pattern is stored.

Measures: The stored data are the patterns of the paths of connections to webpages.

This way, each cube is seen as an array composed by two dimensions: access frequency and length of the path. In addition, each place of the matrix contains a list that represents the stored pattern (See fig. 2).

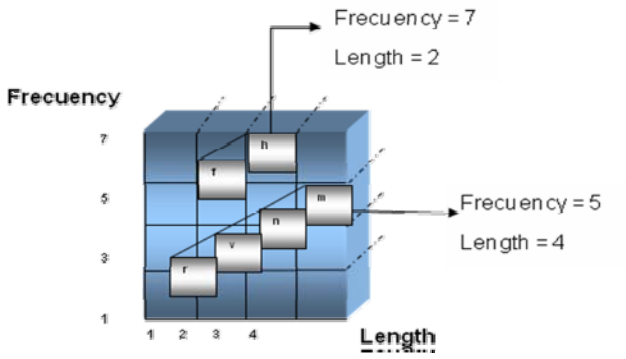


Fig. 2. Cube structure

3.3 Consultations Subsystem

This subsystem generates a graph through the application of the technique of analysis of paths, in which all the cubes built in to previous part are used in the search tasks of the information contained in the system. Two modules form this subsystem:

PATHS ANALYSIS MODULE: This technique allows generating the structure that organizes the relationship among the cubes in which the sequential patterns are stored. Based on this, an indirect graph is built formed by nodes whose arcs have a specific value. The weight of these arcs is calculated based on the number of connections that they have in both cubes. That is, the number of connections contents in a cube C1 that equally are contained in the other cube C2, divided by the total number of connections that exists in the two cubes.

$$\text{Weight} = \frac{\text{Number of common webpages}}{\text{Total webpages of the cubes}}$$

We could use others procedures to calculate the weights of the arcs defined in the literature [9, 11, 13, 14, 15].

In the figure 3, we can observe that the common connections among them are “a” and “c”, then we proceeds to count the number of times that they appear in each one of these cubes. For example:

In cube C1 “a” appears 3 times and “c” appears 2 times.
In cube C2 “a” appears 2 times and “c” appears 1 time.

Therefore, the weight among them is 8/13 = 0,62

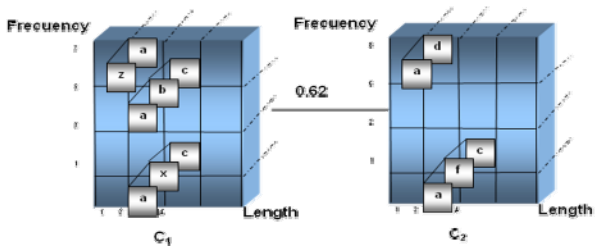


Fig. 3. Weight Computing

MODULE OF SEARCH OF INFORMATION: Once established the structure that contains all the gathered information, and organized it, an algorithm is applied to make searches of information on that structure. For that, we use a search algorithm in width (the breadth-first search algorithm [5, 6, 7], see fig. 4), based on examining all the adjacent nodes to a node, then the adjacent nodes to these, and so forth until to find the searched node or to finish the path of the graph.

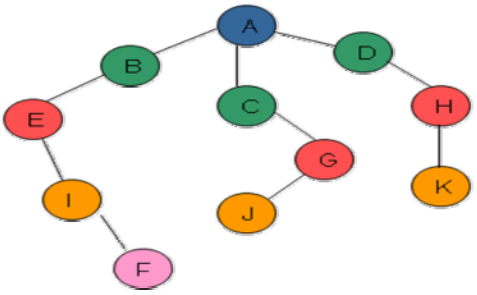


Fig. 4. the breadth-first search algorithm (In this case, the order of route is: A-B-C-D-E-G-H-I-J-K-F)

4 Case of Study

The case of study has like main objective to describe the behaviour of the users that visit to CEIDIS (Coordination of Interactive Studies at the University de los Andes) website. For it, we

have used a subset of the logs files of the webserver (Registrations of January to May of 2008).

4.1 Operation of the System

At the beginning, our Web System need to be configured with the routes of the web server and the log files to be used in our system (see figure 5)

Fig. 5. Initial screen of the web system

The Hybrid System of Web Mining allows carrying out several tasks. One of the most important tasks is the search. The user can choose one of the search options: for keyword or for page (see fig. 6).

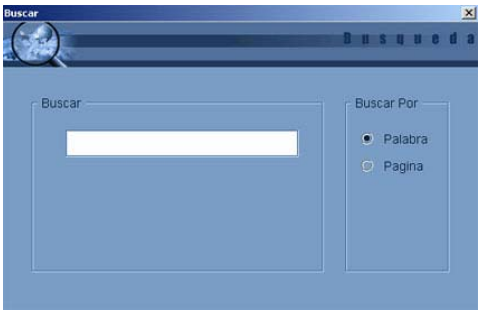


Fig. 6. Search window

4.1.1 Search for keyword

It is based on the consultation of the information stored according to the keyword that describes to each one of the pages, which constitute the website

- If the search is failed, a window is shown where is indicated that the requested information was not found.
- If the result of the search is positive, a list of patterns that have that keyword appears in the inferior part of the main window (see figure 7).

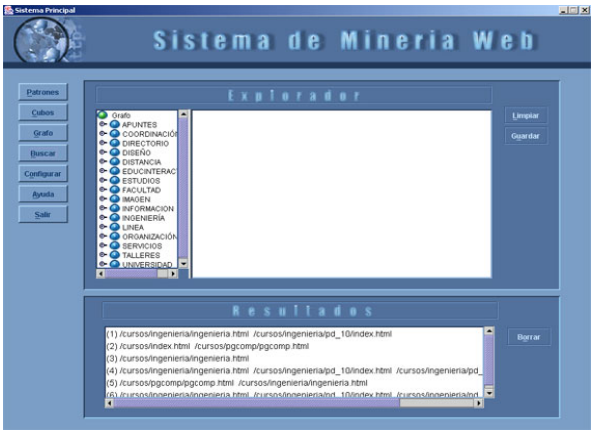


Fig. 7. Results of the search

When pulsing on anyone of these patterns, a window is shown where is analyzed the selected pattern. Three sub-windows compose this window: In the first of them the pattern's analysis is shown, where values of interest are specified (see figure 8). These values represent the keyword of the pattern, the access frequency to it, its length, the use percentage with respect to the rest of the patterns, and the category to which one belongs.

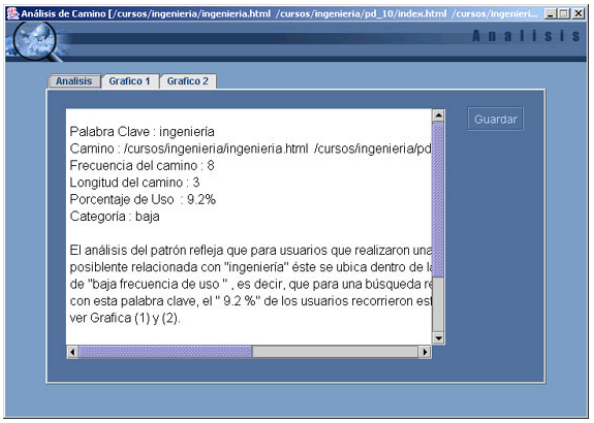


Fig. 8. Analysis window

In the second sub-window a graphic is presented that contains all the patterns found for the search in question, where we can observe the frequency for each one. The selected pattern has a different tonality (blue colour, see figure 9). The graphic of this sub-window is composed by two variables: pattern frequency and pattern. The patterns are organized in descending order.

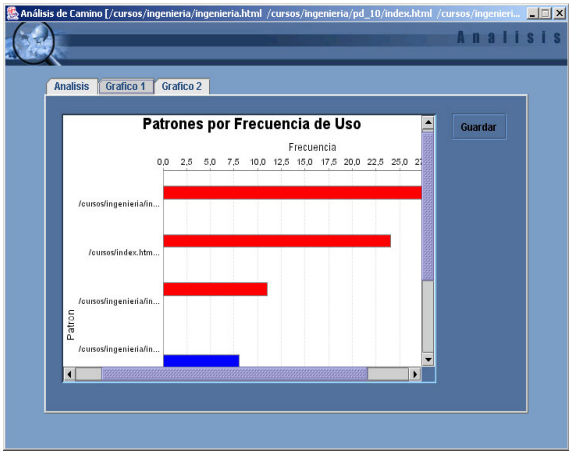


Fig. 9. Frequency graphic

In the third sub-window another graphic is presented, in which the patterns are shown according to the established categories. These categories are product of the evaluation of the frequency of use of these patterns. The categories can be: high, medium and low. Each one of them is represented in the graphic with a different colour (red for high, blue for medium and green for low, see figure 10). With this graphic, we can obtain information like the patterns with more access, the number of patterns that belongs to certain category, differences between groups of patterns that compose a specific category, among other things.

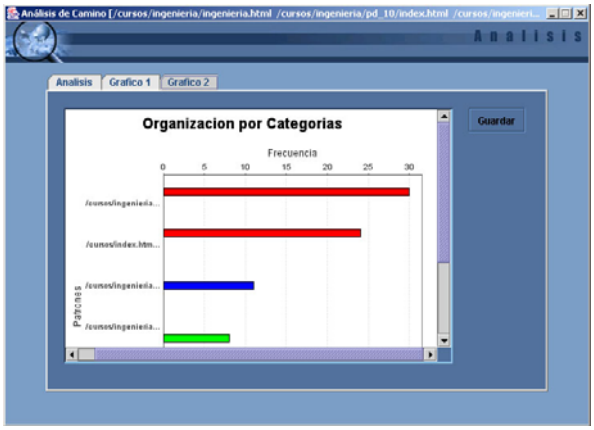


Fig. 10. Results of the search by the keyword: “faculty”

For example, the pattern:

/cursos/index.html /cursos/ingenieria/ingenieria.html
/cursos/ingenieria/pd_10/index.html

has an use frequency of 10, and also it belongs to the category “medium”. It is presented as a result of a search by the keyword “faculty” (see figure 11).

Fig. 11. Results of search for keywords: “faculty”

On the other hand, the pattern:

/cursos/ingenieria/ingenieria.html
/cursos/ingenieria/pd_10/index.html

has an use frequency of 15 and it belongs to the category “high”. It is presented as a result of a search with keyword “engineering” (see figure 12).

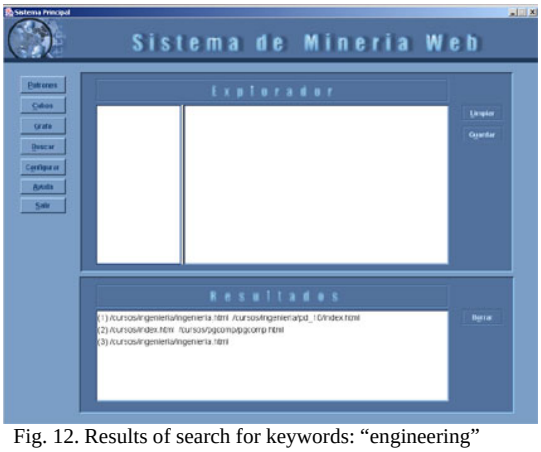


Fig. 12. Results of search for keywords: “engineering”

One can deduce that enough users are using the second pattern because there is some part of interest for the topic of Engineering, while the topic of Faculty is less interesting.

Additionally, we can obtain the list of paths by frequency of utilisation. Figure 13 shows the list of path with a frequency of access of 5.

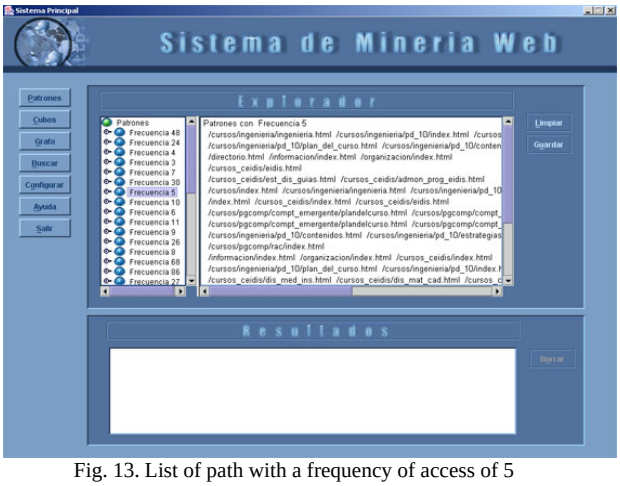


Fig. 13. List of path with a frequency of access of 5

4.1.2 Search for keyword

This search type is focused in the study of the pages that constitute the website. When the user writes the first three letters of the requested page, then appears a list of connections that contain that word. The user can choose one of them (see figure 14).

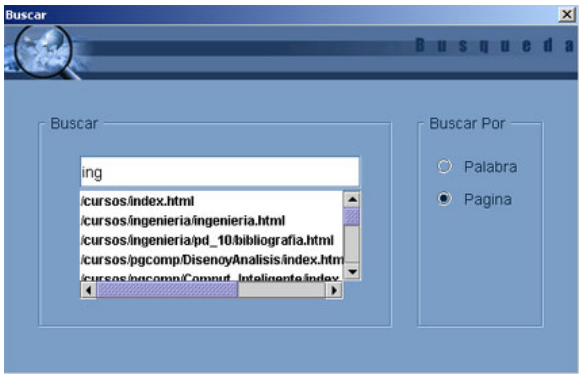


Fig. 14. Page search

Additionally, we can obtain the webpages that compose a given cube (remember that each keyword is described by a cube, see figure 15).

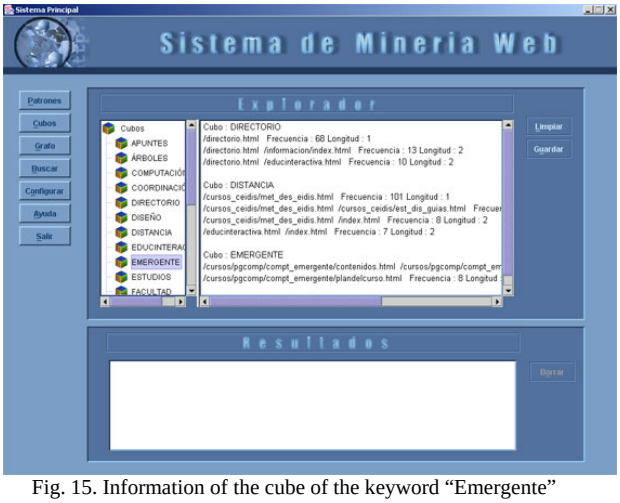


Fig. 15. Information of the cube of the keyword “Emergente”

We can select one of the pages of the list of the fig. 14.

- if the search of the selected page fault, a window is shown where is indicated that the requested information was not found.
- If the search is found, a window is shown in which an analysis of the found page is presented (see fig. 16).

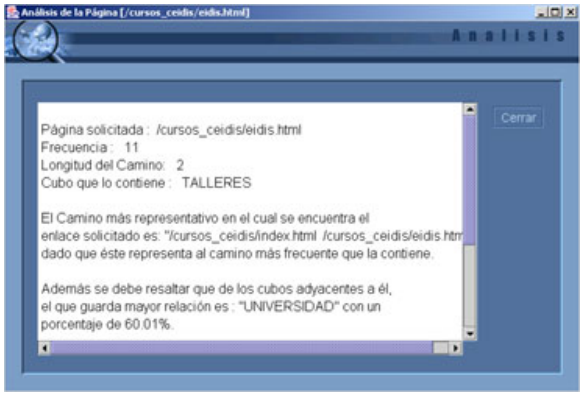


Fig. 16. Information of the Page found

In this window the next information is presented: the frequency of use of the requested page, length of the most frequent path to arrive to it, and the cube that contains it. Consequently, the name of the cube is the keyword of this pattern. Some analyses of interest are also presented like the relationship that exists among the cube that contains this page and the adjacent cubes to it

The last graphic that can obtain with our system is the interconnection graph of the cubes, where we can see the weights of the arcs, that is the relationships between them (see fig. 17).

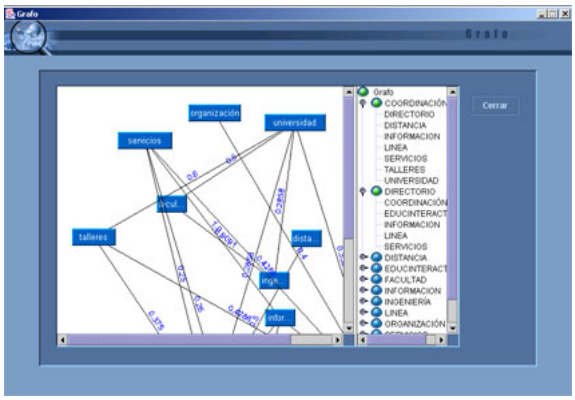


Fig. 17. Interconnection graph of the cubes

4.2 Result Analysis

Analyzing the patterns obtained with the system, certain aspects characteristic of the area of web mining can be observed. For example: For a search by keyword, in which the user

requests information on “Distances”, these are the results:

- /cursos_ceidis/met_des_eidis.html
- /cursos_ceidis/met_des_eidis.html /cursos_ceidis/est_dis_guias.html
- /cursos_ceidis/met_des_eidis.html /index.html
- /educinteractiva.html /index.html

It is important to indicate that these paths are in descending order according to the frequency of use of the users of the Website. This list of paths is shown in the main window of the system, in the area of Results

But, this result is not only the list of webpages with the keyword “distance”, rather we also can carry out an analysis of each one of the paths. For example, if we take the third path, a window is shown in which the following information is observed:

Keyword: distances
Path: /cursos_ceidis/met_des_eidis.html
Frequency of the Path: 101
Length of the path: 1
Use percentage: 81,16%
Category: High

- The analysis of the path reflects that for a search of the word “distances”, 81,46% of the users use this path. This way, this path is located inside the category “high.” That is, it is the webpage with the content with more relationship with the keyword “distances” in this site.
- Another observation is given by the number of accesses that the users made to arrive to this page. Because the length of the path is 1, this indicates that the information that the user searched was found directly in this page. In this way, one can affirm that the user accessed directly to the page and not for the main page of the site; this allows to deduce that the user had a previous knowledge on the site.

Also, for the case of a search for page, in which the user requests the page:

/cursos/ingenieria/ingenieria.html,

these are the results:

Requested page: /cursos/ingenieria/ingenieria.html
Frequency: 30
Length of the path: 2
Cube that contains it: ENGINEERING

Of the adjacent cubes to ENGINEERING, the one that has bigger relationship is: "ABILITY", with a percentage of 78.95%. This indicates that bigger relationship exists among the information contained in the cube ENGINEERING and the cube ABILITY. Also, 78,95% means the number of common connections that exist among these cubes; that is, of each 10 connections 8 approximately are in both cubes. This allows deducing that exist points in common among the users that visit the pages that compose these cubes. For example, students with a profile of preference by the contents of these pages.

Other values of interest can be obtained. For example, analyzing the path:

/cursos/index.html /cursos/ingenieria/ingenieria.html

this has a frequency of 14 . This indicates that users that access to the webpage /cursos/ingenieria/ingenieria.html, they access first to /cursos/index.html. This way, one can affirm that some information in /cursos/index.html addresses the users to cursos/ingenieria/ingenieria.html.

Particularly, the webpage /cursos/index.html presents the information on the different courses that this website offers. However, when observing all the found paths, the biggest access frequency is presented in the engineering courses; this indicates that the biggest number of users that access this website are interested in studies in the engineering area. The courses of the other areas and the way like they are presented in /cursos/index.html should be revised, so that the interest of the users in visiting them increases.

We can see that a lot of the paths found have a length between 1 and 2, which indicates that the users abandoned the site after consulting two or less pages. This indicates that most of the users don't navigate more than two pages in this site. We need to be sure that the most important information is contained inside the main pages of this site.

This information is of vital importance for the administrator of a website. For it, the knowledge obtained with this application can be used in the personalization of contents (generation of contents, guided navigation, among other) like in the restructuring of the website [ref].

Starting from data of a file that simply represent the registration of users' visits to a website, we have obtained very valuable information. The main aspects studied are the structure of the site and the information contained in it.

4 Conclusion

In this work has been investigated the use of techniques of web mining. The work proposes a System of Web Mining that is a hybrid of several techniques of this area. This system is a visualization and analysis of information tool based on the log file of the users in a website. The proposed system is based on the integration of the next techniques of web mining: sequential patterns, cubes and analysis of paths, with the objective of finding paths frequently used by users when use Internet. Access patterns, tendencies and groups are also identified among users that visit a specific website. The most significant contributions are:

- The proposed System of Web Mining is a hybrid system that proposes to mix different techniques of web Mining, exploiting the advantages of each one of them.
- The techniques are used in such a way of organizing the information on the use of the web efficiently, in such a way of facilitating the access later, exploiting the information over the discovered knowledge.
- One of the main tasks carried out in the system consists on to find and to analyze patterns of the users' access in a website. This allows knowing the preferences of the users regarding the contents of the pages that constitute this site.
- The system facilitates the analysis of the obtained information, in such way that we can discover tendencies of behavior, of vital importance for the increment of users' visits to the website, as well as improve the web service offered.

However, the system presents certain aspects that should be analyzed and improved in later studies. One of them is the selection of the keywords that represent a pattern, because situations can be given in which all the words have the same repetition number and in the current version we choose one randomly. Another one is about the similarity relationships among cubes. In the current version we consider the common webpages in each cube.

One of the future investigations is the development of intelligent tools that can help in the interpretation of the discovered knowledge, for the analysis of the behavior and the users' tendencies. We could use techniques of the area of artificial intelligence, like the artificial neural networks, to study the discovered patterns for our system, for the automatic analysis of the obtained information, which would allow to generate new knowledge.

References:

[1] Lozano A., Gómez A., Sosa E. Selection of Ontologies for the Semantic Web, *Lecture Notes in Computer Science*, Vol. 2722, 2003. pp. 413-416.

[2] Castells P. La web semántica. In *Sistemas Interactivos y Colaborativos en la Web* (Eds Bravo C., Redondo M.). Ediciones de la Universidad de Castilla, 2005, pp. 195-212.

[3] Masand, B. Zaiane, O. Srivastava, J. Spiliopoulou, M. Web Mining for usage Patterns and Profiles, *ACM SIGKDD Explorations Newsletter*, Vol. 4, No. 2, 2002, pp. 125-127.

[4] Aguilar J., Altamiranda, J. Minería de Datos en la Web usando Computación Evolutiva, In *Ingeniería de Software en la Década del 2000s*. (Ed. Brisaboa N.), AECI, 2006 pp. 153-168.

[5] Aguilar J., Leiss E., Callaos N. *Introduction to Web Computing*. Intl. Institute of Informatics & Systems, 2004.

[6] Aguilar J., Altamiranda J. "Conceptos sobre Minería Web". *Revista Gerencia Tecnológica Informática*, Vol 3, No. 7, 2004, pp. 71-77.

[7] Bagües M., Bermúdez J., Illarramendi A., Tablado A., Goñi A. Semantic interoperation among data systems at a communication level. *Journal of Data Semantics*. Vol. 3870, 2006, pp. 1-24.

[8] Wasniowski, R. Data Mining Support for Intrusion Detection and Prevention, *WSEAS Transactions on Computer Research*, Vol. 1, No. 2, 2006, pp. 355-339.

[9] Castellano, G. Fanelli, A. Torsello, M. Understanding Visitor Behaviors from Web Log Data, *WSEAS Transactions on Computer Research*, Vol. 2, No. 2, 2007, pp. 277-284

[10] Zhang, W. Wang, Y. Towards Building a Semantic Grid for E-Government Applications, *WSEAS Transactions on Computer Research*, Vol. 3, No. 1, 2008, pp. 273-282.

[11], Taowei Wang, Yibo Ren, Research on Personalized Recommendation Based on Web Usage Mining Using Collaborative Filtering Technique, *WSEAS Transactions on Information Science and Applications*, Vol. 6, No. 1, 2009, pp. 62-72.

[12] Cheng, Ch. Huan, S. Chuang, M. A Study on the Applications of Data Mining Techniques to Enhance Customer Lifetime Value, *WSEAS Transactions on Information Science and Applications*, Vol. 6, No. 1, 2009, pp. 319-328.

[13] Tang, J. The Considerations of the Web Page Design, *WSEAS Transactions on Information Science and Applications*, Vol. 6, No. 4, 2009, pp.637-646.

[14] El-Bakry, H. Mastorakis, N. Fast Information Retrieval from Web Pages, *WSEAS Transactions on Information Science and Applications*, Vol. 6, No. 6, 2009, pp 1018-1036.

[15] Herceg, T. Jakovic, B. Markovic, M. Analysis of Retailer Web Sites with Microeconomic Interpretation, *WSEAS Transactions on Information Science and Applications*, Vol. 5, No. 4, 2008, pp. 557-568.

[16] Borges, A. Gil, R. Corniel, M. Contreras, L. Borges, R. Towards a Study Opportunities Recommender System in Ontological Principles-based on Semantic Web Environment, *WSEAS Transactions on Computers*, Vol. 8, No. 2, 2009, pp. 279-291.