User Science Indicators in the Web Context and Co-usage Analysis

Xavier Polanco, Ivana Roche, Dominique Besagni
Unité de Recherche et Innovation, Institut de l'Information Scientifique et Technique,
Centre National de la Recherche Scientifique,
2 allée du Parc de Brabois – 54514 Vandoeuvre-lès-Nancy
{polanco,roche,besagni}@inist.fr

We present a new kind of statistical analysis of science and technical information (STI) in the Web context. We propose a battery of indicators about Web users, used bibliographic records and e-commercial transactions. In addition, we introduce two Web usage factors and we give an overview of the co-usage analysis. For these tasks, we present a computer-based system, called Miri@d, which produces descriptive statistical information about Web users' searching behaviour, and what is effectively used from a free-access digital bibliographical database.

1. Introduction

There are two traditions about the analysis of the Web: one developed by people coming from documentation and the other by computer scientists. The first was developed in the field of information science under the appellations of "webometrics" ((Almid & Ingwersen, 1997), or "cybermetrics" (cf. http://www.cindoc.csic.es/cybermetrics) while seeking to extend the informetric techniques to the analysis of the Web (Björneborn & Ingwersen, 2001; Ingwersen & Björneborn, 2004). The second arose in the field of computer science while seeking to extend the data mining techniques to Web analysis under the appellation of "Web mining" (Chakrabarti, 2003) and according to three main categories: "Web structure mining", "Web content mining", and "Web usage mining" (Kosala & Blockeel, 2000). We work at the border of these two traditions: we consider informetrics from the point of view of computer-based technologies. The Web represents a new environment for the quantitative studies of science, and a new family of computer-based science indicators can be developed. This article deals with a system able to produce descriptive bibliometric statistics, and statistical information on Web users' behaviour.

In comparison with traditional bibliometric studies and in addition to statistics on scientific bibliographical data, we can now analyse the queries, i.e. what the users wish to obtain, and how they express their requests, and their respective top-level domain (TLD, e.g. ".fr" or ".com"), as well as the economic transactions, in which users become customers ordering copies of the retrieved documents. Thus, bibliometrics is encapsulated in the Web usage analysis.

The article is organized as follows: Section 2 deals with the presentation of the Miri@d server, and the statistical information that Miri@d is able to produce. Section 3 describes the co-usage analysis and its application on Web user data coming from the Miri@d server.

2. A Web statistical tool

The system, called Miri@d, is conceived as a server of statistical data which are carried out beforehand, and as an interactive server for online statistical work. In fact, the Miri@d server

produces descriptive statistical data about both what is effectively used from a source of information, which is available as a free-access digital bibliographical database, and the searching behaviour of Web users trying to fill an information need with a scientific multi-disciplinary source of information put at their disposal. The results will be made available to analysts, who can use this descriptive statistical information as raw data for theirs indicator design tasks, and as input for multivariate data analysis, clustering analysis, and mapping. Managers also can exploit the results in order to improve management and decision-making.

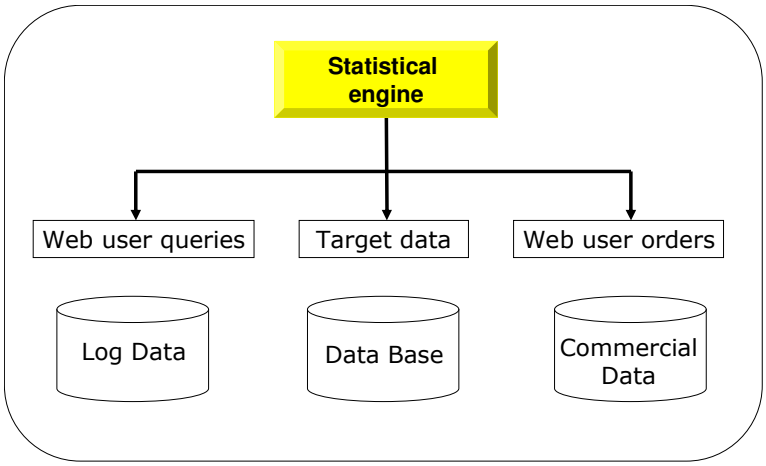

Figure 1: The model

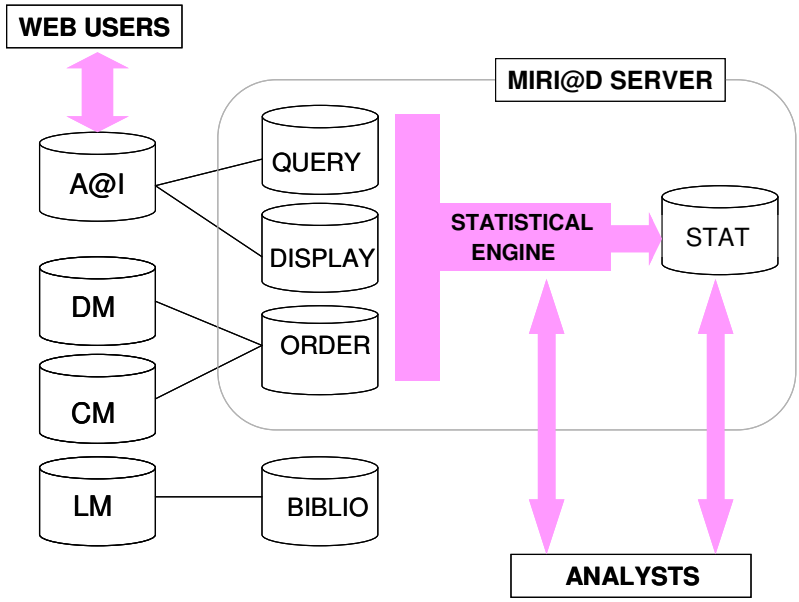

Figure 2: The server

Figure 1 represents what we call the model. The model is general in the sense that it is not limited to the particular characteristics represented. It is significant to see that the model implies three families of data which can be exploited: on the one hand, log-files data and on the other hand, bibliographic data and commercial data. From the point of view of scientific information, the bibliographic database can be replaced by any other type of database, at least theoretically. On the level of its concept, the model is more general than the one specified by the data sources it is using at the moment.

Figure 2 represents the server structure, which consists of a set of external resources providing the statistical raw data, and a set of database internal to server. The resources from which Miri@d receives data are the following: A@I is the server Article@INIST that provides the bibliographic database and log-files, DM is the document delivery management system, CM is the customer management system, and LM is the library management system. The Miri@d server own databases are QUERY: data related to Web users' queries, DISPLAY: data related to displayed bibliographical records, ORDER: data related to ordered documents, BIBLIO: bibliographical records, STAT: statistical indicators calculated beforehand on the data stored in the other Miri@d databases. The statistics are calculated off-line with a periodicity of one day, one week, one month or one year.

2.1 New Science indicators in the Web context

Table 1 show the set of statistical data that constitute the output of the Miri@d statistical engine, and that analyst can access or interactively produce through the Miri@d server.

Web User and Usage	**Scientific Publication**	**E-commerce**
Number of queries	Number of displayed records	Number of ordered documents
Distribution by: - users' TLD - users' country - query date - title word (in query) - author (in query) - keyword (in query) - number of obtained records - number of displayed records	Distribution by: - users' TLD - users' country - publication year - document type - author - journal - publishing country - language	Distribution by: - customers' country - customers' activity - publication year - author - journal - publishing country - language

Table 1 – Descriptive statistical information

To illustrate the STAT database, we provide some examples of statistical information as the distribution of web users' searches by users' countries during the year 2002. From 135 countries, those constituting the European Union are at the origin of 872,510 searches corresponding to more than 84%, and around 92% of these searches come from France. Concerning displayed records, if we consider the same year, the total number of displayed journals and records in this period are equal to respectively 16,797 and 391,693. Concerning ordered documents, the total of ordered documents, 71,208 for the year 2002, is split by type of customers' activity as follows: 54% commercial firms, 28% research institutions, 9% higher education and 9% others. The typology of users' activities adopted by Miri@d is the one used by INIST's Customer Management System.

To illustrate the QUERY database utilization, we selected a set of queries asking for the word "polymer" in a journal or article title. A set of 2,511 queries is obtained and their distribution

by users' country permits to observe that they come from 28 countries. France is the country producing the greatest number of queries with 82%. To illustrate the DISPLAY database utilization, we use a set of 917 records displayed by users and linked to the above mentioned set of queries about polymers. The distribution of these records by user countries shows that users coming from eighteen countries have viewed these records. The country with the greatest number of displayed records is France with 79%. As for the ORDER database, 4,513 ordered documents coming from 57 journals were related to polymers. The distribution of the 4,513 ordered documents by customers' country shows that documents have been ordered in 15 countries, with 86% of orders coming from France. The most represented customer activity is "Commercial firm" with 46% of orders, but "Research institutions" is not far behind with 45%. The only other significant value belongs to "Higher education" which represents 8%.

From these data, we can note that this multi-disciplinary database is mostly used in Europe, especially in France.

2.2 Web usage factors

In addition, we introduce two user indicators dealing with Web users' information retrieval and Web customers' orders. The first is a Web usability factor, and the other is a Web customer order factor. These factors can be considered for evaluating online information sources by the observation of the information displayed by users, and the documents ordered by user-customers. A situation is the number of times an information source is used or displayed by online users. This is a well known situation in information retrieval. The other is the number of times an information source is ordered; in this case we are in face of e-commerce transactions.

The **Web Usability Factor (WUF)** is the proportion of articles of a journal displayed by Web users in a period of time from t_0 to t_1 by the total of articles published in this journal and stored until t_1.

$$WUF_{JT} = \frac{\sum_{i=t_0,t_1} \sum_{\substack{m=JT \\ \forall n}} dr_i(m,n)}{JT_{t_1}}$$ With: dr = displayed records, JT = journal title, n = publication year, JT_{t_1} = number of journal title articles, all publication years, stored until t_1. For example, bibliographical records concerning the "Journal of applied polymer science" were displayed 2,858 times in 2002, which gives a WUF of 0.18.

When the WUF index is calculated on a yearly basis it becomes possible to study journal titles obsolescence.

$$WUF(PY)_{JT} = \frac{\sum_{i=t_0,t_1} \sum_{\substack{m=JT \\ n=PY}} dr_i(m,n)}{JT_{t_1}(PY)}$$ With: dr = displayed records, $[t_0,t_1]$ = period of time, JT = journal title, PY = publication year considered, $JT_{t_1}(PY)$ = number of journal title articles in publication year PY, stored until t_1.

The **Customer Order Factor (COF)** is the proportion of articles of a journal ordered by Web customers in a period of time from t_0 to t_1 by the total number of articles published in this journal and stored until t_1.

$$COF_{JT} = \frac{\sum_{i=t_0,t_1} \sum_{\substack{m=JT \\ \forall n}} ord_i(m,n)}{JT_{t_1}}$$ With: *ord* = ordered documents, *JT* = journal title, *N* = publication year, JT_{t_1} = number of journal title articles, all publication years, stored until t_1. For example, articles from the "Journal of applied polymer science" were ordered 366 times in 2002, which gives a COF of 0.022.

When the COF index is calculated on a yearly basis it becomes possible to study journal titles obsolescence.

$$COF(PY)_{JT} = \frac{\sum_{i=t_0,t_1} \sum_{\substack{m=JT \\ n=PY}} ord_i(m,n)}{JT_{t_1}(PY)}$$ With: *ord* = ordered documents, $[t_0, t_1]$ = period of time, *JT* = journal title, *PY* = publication year considered, $JT_{t_1}(PY)$ = number of journal title articles in publication year PY, stored until t_1.

3 Co-usage analysis

Co-usage analysis belongs to the same family as co-citation and co-word analyses. We expose in this section the formal characteristics of co-usage analysis, which obeys the co-occurrence framework for clustering and mapping the Web usage. We use it for identifying and visualising usage centres of interest on certain research foci, or problem areas. Co-usage analysis will be applied on the Web users' data produced by Miri@d. Actually, co-usage analysis is not yet included in Miri@d. Another Web server called VISA makes it possible for analysts to reach the co-usage analysis results.

3.1 Co-usage concept

Documents d_i and d_j are related by usage coupling, when a user u_i refers in his or her query to documents d_i and d_j. If $u_i \rightarrow d_i$ and ui $\rightarrow d_j$, then d_i and d_j are associated d_{ij} by u_i, with $i = 1, 2\ldots m$. Conversely, users u_i and u_j are related as a pair of users, u_{ij}, when they refer in their queries to a same document d_i. If $u_i \rightarrow d_i$ and $u_j \rightarrow d_i$, then u_i and u_j are associated by d_i, with $i = 1, 2\ldots n$. A given document or bibliographic data is represented by an usage attribute vector of the form: $d_i = (u_{i1}, u_{i2}, \ldots, u_{im})$, where u_{ij} is an identifier representing the *jth* user reference of the document *i*. Vector similarity operation is performed using user references similarities. A given user can be represented by a record (bibliographic data) attribute vector of the form: $u_i = (d_{i1}, d_{i2}, \ldots, d_{in})$, where d_{ij} is an identifier representing the *jth* document referred by user *i* and displayed or ordered by the same the user *i*. Vector similarity can be performed using displayed documents similarities. Thus, each document d_i can be described by the set of co-users, u_{ij}; $u_i = \left[d_{ij} \right]_{i=1,m;\, j=1,n}$; with $d_{ij} \in [0,1]$; and *m* = number of users, *n* = number of documents. In the same way, each user can be described by a set of co-documents d_{ij}, $u_i = \left[d_{ij} \right]_{i=1,m;\, j=1,n}$; with $d_{ij} \in [0,1]$; and *m* = number of users, *n* = number of documents.

An association coefficient is used for normalising the co-occurrences in *D* and *U* matrices. There are several methods of calculating an association coefficient. We use the so called "equivalence coefficient" largely used in co-word analysis and which it is defined as:

$$E_{ij} = \frac{\left[C_{(i,j)}\right]^2}{c_{(i)} \times c_{(j)}}$$

C(i,j) is the total number of co-occurrences of users *i* and *j* or documents *i* and *j*, *c(i)* is the total number of occurrences of item *i.*. This association coefficient is analogous to the well known Dice, Jaccard, Ochiai, or Salton coefficients. As results of the application of this association coefficient on user couples, we obtain a real valued matrix, denoted by *A*, and composed of the association coefficients values. The association matrix *A* gives a normalized measure of the strength of associations between users *AU*, and between documents *AD*. We apply a single linkage hierarchical agglomerative clustering method to these matrices.

The algorithm we use is an adaptation of the standard bottom-up single-link clustering in accordance with readability criteria on the size of the cluster, which is defined as the minimum and maximum number of items belonging to the cluster, and on the maximum number of associations constructing the cluster. As a consequence of this clustering process, two kinds of associations and items forming the clusters are generated: one is said to be internal or intra-cluster and the other external or inter-cluster. If both elements of a given pair belong to the same cluster, the association between the items is considered as an internal association of that cluster. If they belong to two different clusters, the association is considered as an external association. The items involved in internal associations of a cluster are called internal items. The number of internal items defines the size of a cluster. Those items rejected during clustering because they do not meet the "maximum cluster size" criterion are recorded as external items.

3.2 Co-usage analysis application

A detailed interpretation of the results is out of question in this article. We limit here to expose the results that the co-usage analysis provides to analysts who would perform the interpretation task. However, we will underline especially the means of analysis provided to assist the work of interpretation and decision. From Miri@d's ORDER database, a set of documents dealing with polymers was collected. These data consist of scientific articles that have been ordered by customers after consultation of the Article@INIST repository. The size of the ORDER database is the 202,391 orders through two years between 2001 and 2003. For this application, the dataset is equal to 3,914 documents published by 57 journals and ordered by 410 customers.

Since each document d_i can be described by the set of customers u_{ij}, and reversely each user u_i can be described by the set of documents d_{ij}. Two co-usage analyses have been done: one on the ordered documents and the other on the user-customers (or authors of the orders.) In the next sections, we describe these applications at the levels of clusters (see Figures 3 and 4) and maps (see Figures 5 and 6)

3.2.1 Co-usage analysis of the ordered documents

24 clusters were obtained from the ordered-document co-occurrence matrix $U(d_{ij})$ $i = 1,\ldots,m$ and $j = 1,\ldots,n$; m = number of user-costumers = 410, n = number of ordered-documents = 3,914. They are constituted by ordered-documents dealing with subjects interesting sub-sets of users. The documents are represented by codes. Figure 3 shows the graph structure of a cluster. The map is represented in Figure 5.

3.2.2 Co-usage analyse of the user-customers

17 clusters were obtained from the user-costumers co-occurrence matrix $D(u_{ij})$ $i = 1,\ldots,n$ and $j = 1,\ldots,m$; m = number of user-customers = 410, n = number of ordered-documents = 3914. These clusters are then constituted by user-customers who are interested by sub-sets of ordered-documents in which the knowledge interesting these user-customers is embedding. Figure 4 shows an example of the graph of a user-customer cluster. To preserve confidentiality the customers are represented only by their country and activity sector codes. The map of the user-customer clusters is presented in Figure 6.

3.2.3 Analytical tools: clusters and maps.

Clusters and maps constitute analytical tools. A cluster is composed of items that are called internal items. The internal item with the maximal weight value $w_{Cl}(a)$ is automatically chosen to be the cluster label. Let be a an internal or external item (document or user) of the cluster Cl, then the weight is defined as follows:

$$\underset{a=1,m_{Cl}}{w_{Cl}(a)} = \frac{k_{Cl}(a)}{n_{Clin} + n_{Clex}} \text{ with: } 0 < k_{Cl}(a) \le n_{Clin} + n_{Clex} \text{ And: } 0 < w_{Cl}(a) \le 1$$

Symbol sense: m_{Cl} is the number of its internal and external items, n_{Clin} is the number of its internal associations, n_{Clex} is the number of its external associations, $k_{Cl}(a)$ is the number of occurrences of internal or external item a $(a = 1, m_{Cl})$ in the internal or external associations of Cl.

The clusters are also composed of associations between these items which are also called internal associations, to distinguish them from external associations which link a cluster with other clusters. The internal and external associations are weighted relations according to E_{ij} association coefficient (see section 3.1). The internal associations between the component items of a cluster are represented by the graphs in Figures 3 and 4. On the other hand, the links between clusters are shown on the maps of the Figures 5 and 6.

In Figures 3 and 4 two classes of internal items are defined: leaving and entering. This allows us to know by which item this cluster is connected to another cluster, and at the same time which are the internal items by which the cluster receives links starting in another cluster. According to this angle of analysis, the clusters can be analysed as insulated, receivers or senders within the network of clusters.

On the maps, clusters are placed in function of their structural properties, that is centrality (X-axis) and density (Y-axis). The density is equal to the average of the values E_{ij} of internal associations of the cluster. The average of the values E_{ij} of external associations of a cluster is called centrality. Each labelled point in the maps represents a cluster. Following the standard co-word analysis, four types of clusters can be distinguished: clusters with high density and centrality (type 1), with low density and high centrality (type 2), with high density and low centrality (type 3), and clusters with low values on both axes (type 4). The map (called strategic diagram) is then analysed in these terms.

On the other hand, clusters appear within a network. We can visualise in the maps the network of inter-cluster relations. Thus, the co-usage analysis can be extended as a network analysis using graph theory, and then we can interpret clusters and their network in this other analytical framework (Polanco, 2005). What is significant here is to see that we have two possibilities of analysis, one founded in the tradition of the representation of the clusters in a two-dimensional plan, the other according to the network analysis and the theory of graphs.

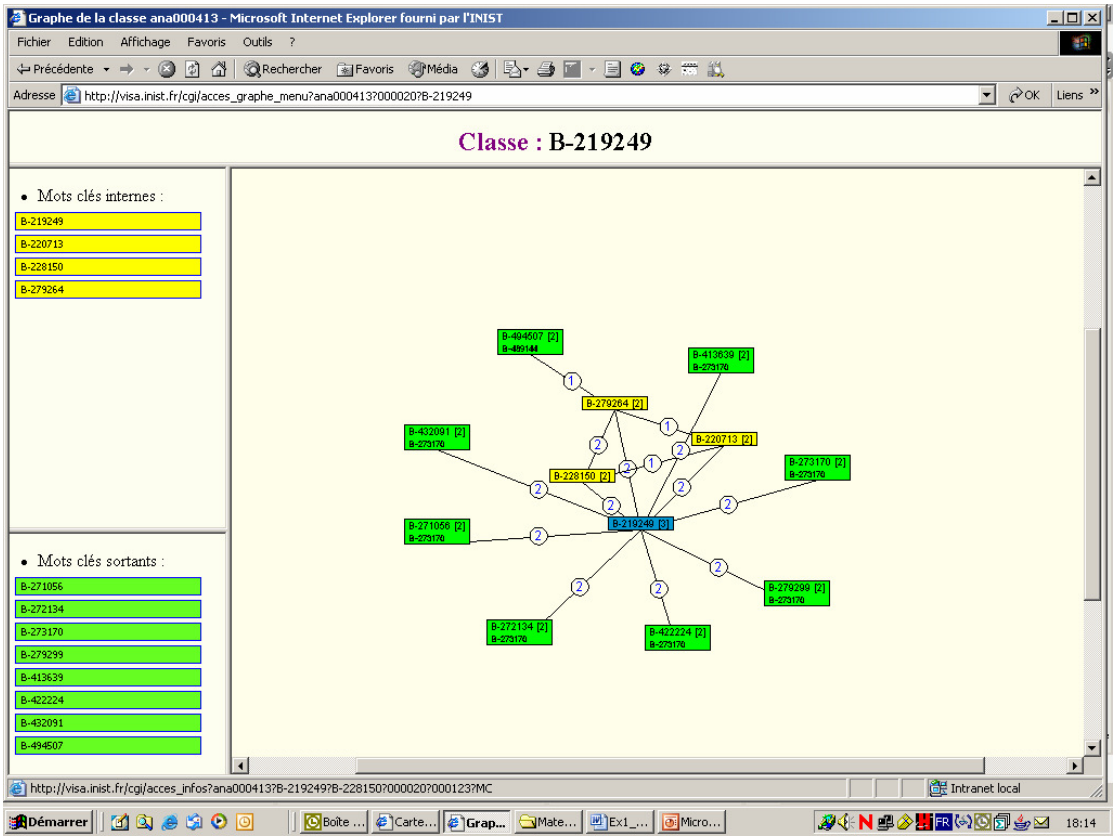

Figure 3: Cluster graph labelled by B-219249 ordered document

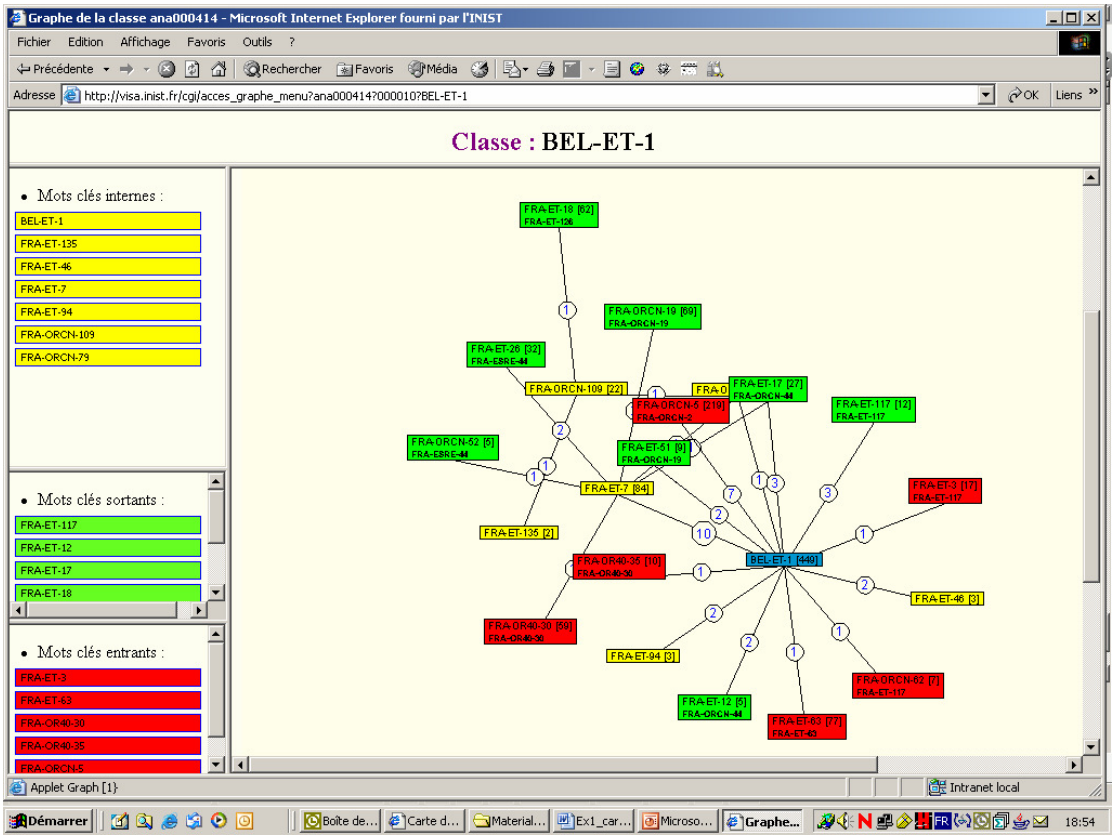

Figure 4: Cluster graph labelled by BEL-ET-1 user-customer

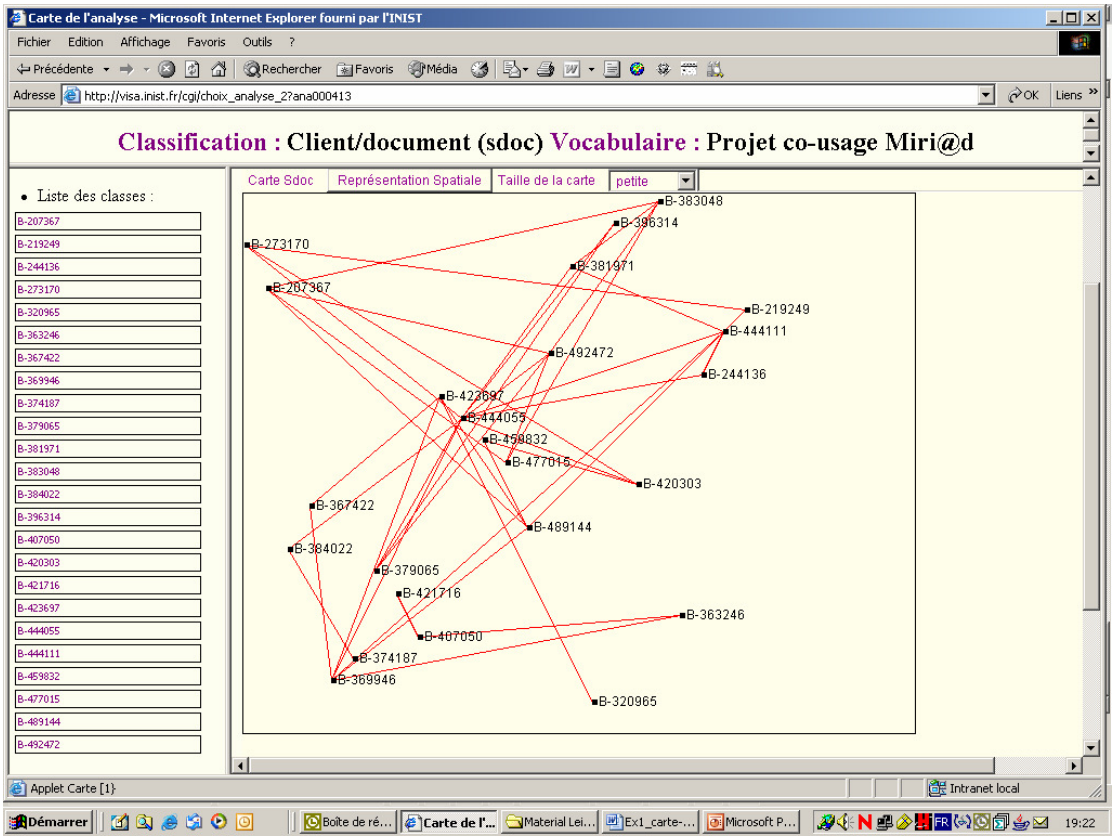

Figure 5: Map 1 Ordered document clusters

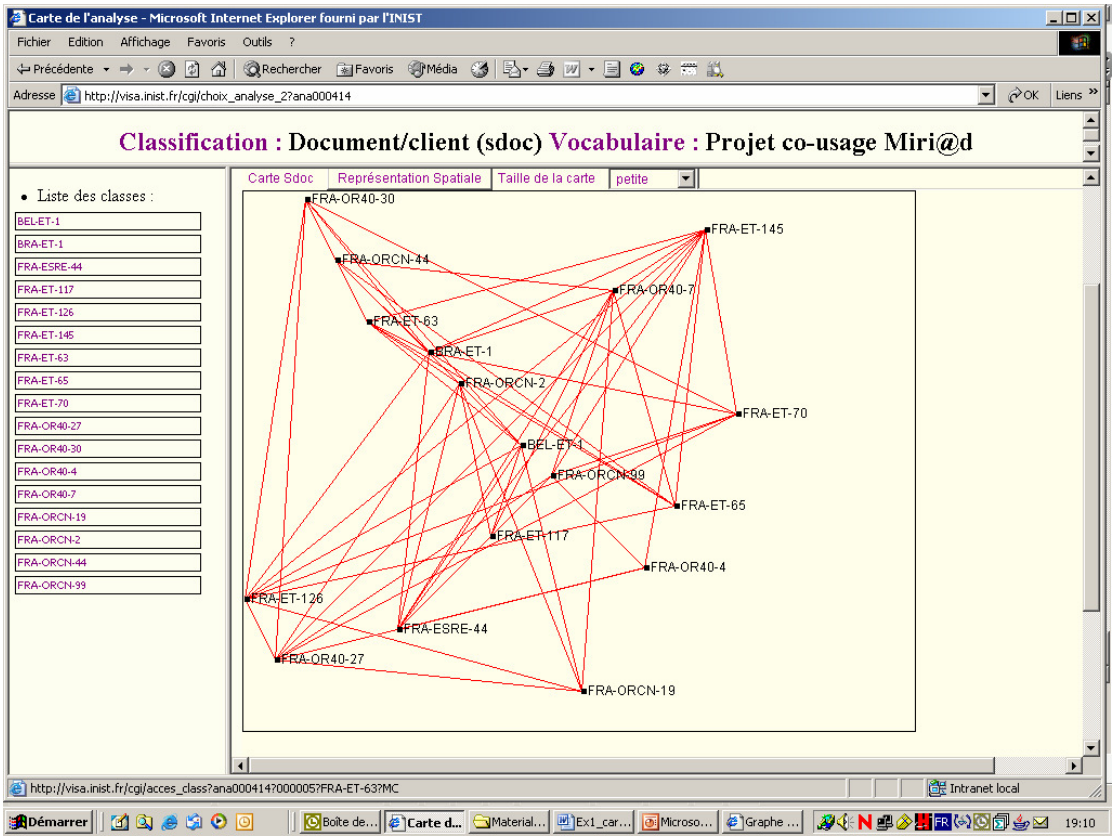

Figure 6: Map 2 User-customer clusters

4. Conclusion

In this article, we presented a Web statistical tool, the Miri@d server, for quantitative analysis of the scientific information. This is conceived to produce descriptive statistical data about what is effectively used (in a free access digital database storing scientific and technical information), and the Web users searching behaviour. We showed that the system is conceived as a server of statistical data which are carried out beforehand, and as an interactive server for personal online statistical work. Two Web usage factors were also proposed.

At the same time, we proposed the co-usage analysis as a generalisation of the co-word analysis, and we remark that the same approach can be used at the level of both the Web content analysis and the Web structure analysis (see Polanco et al., 2001). The application we presented, illustrates the type of information that co-usage provides and the way this information can be analysed. As we said, co-usage analysis belongs to the same family as co-citation and co-word analyses.

We tried to show how today with the Web new ways emerging for analysing science through its publications and their usages. It is the existence of log-files that provides the technical opportunity for introducing the analysis of user behaviours and proposing then the co-usage analysis.

Acknowledgment: The design and development of the Miri@d server, as well as the indicators engineering, and also the co-usage analysis proposition, have been realized thanks to the European project IST-1999-20350 - Fifth Research and Development Framework Plan of the European Union, during 2000-2003. Project acronym EICSTES, and project full title "European Indicators, Cyberspace, and the Science – Technology – Economy System."

References

Almind T.C. and Ingwersen P. (1997) Informetric analyses on the World Wide Web: methodological approaches to "webometrics", *Journal of Documentation*, vol. 53, No 4, p. 404-426.

Björneborn L. and Ingwersen P. (2001) Perspectives of webometrics, *Scientometrics*, vol. 50, No 1, p. 65-82.

Callon M. and Courtial J-P (1997) Using scientometrics for evaluation, in M. Callon, Ph. Larédo, Ph. Mustar (editors) *The Strategic Management of Research and Technology*. Paris, Economica International, chapter 10, p. 165-219.

Chakrabarti S. (2003) *Mining the Web*, Morgan Kaufmann Publishers.

Courtial J-P (1990) Introduction à la scientométrie. Paris, Anthropos.

Ingwersen P. and Björneborn L. (2004) Methodological issues of webometric studies, in *Handbook of Quantitative Science and Technology Research*. Edited by H.F. Moed, W. Glänzel and U. Schmoch. Kluwer Academic Publishers, chapter 15, p. 339-369.

Kosala R. and Blockeel H. (2000) Web Mining Research: A Survey, *SIGKDD Explorations*, vol. 2, No 1, p. 1-15; available at: http://portal.acm.org/

Polanco X. (2005) Modelling co-word clusters as graphs (unpublished preprint).

Polanco X., Boudourides M. A., Besagni D., Roche I. (2001) Clustering and Mapping European University Web Sites Sample for Displaying Associations and Visualizing Networks, in Pre-proceedings *New Techniques and Technologies for Statistics* (NTTS&ETK), 18-22 June, Crete, vol. 2, p. 941-944.

Salton G. (1989) *Automatic Text Processing*. Reading, Mass., Addison-Wesley Publishing.