Implicit Measures of User Interests through Country and Predicting Users' Future Requests in WWW

Pragya Rajput
M. Tech. Scholar, Dept. Of CSE,
Oriental College of Technology
Bhopal, India

Joy Bhattacharjee
Asst. Prof., Dept. of CSE
Oriental College of Technology
Bhopal, India

Roopali Soni
Asst. Prof., Dept. of CSE
Oriental College of Technology
Bhopal, India

ABSTRACT
WWW has become the center of attraction for business transactions and hence e-commerce due to its ease of use and speed. its ability from tracking the browsing behaviour of any user to even the mouse clicks of any individual has brought the vendor and end customer closer than ever before.WWW has made it possible for vendors to advertize their products i.e. they are personalizing their product messages for individual customers at a very large scale such a phenomenon is termed as mass communication such a utility is not only applicable for e-commerce but also such personalization is aiding several web browsing activities. Any action that tailors the web experience to any individual or several users is termed as web personalization. Web personalization is that the method of customizing an internet website to the wants of specific users, taking advantage of the information noninheritable from the analysis of the user's guidance behaviour(Weblog data) in correlation with alternative data collected within the web context, namely, structure, content and user profile information. The domain of web personalization has gained importance both in the area of research and commerce. In this paper we proposed a framework of web log mining to implicit measures of user interests through Country and predicting users' future requests in WWW.

KEYWORDS
Web Log, Personalization, Web Usage Mining, Preprocessing, IP-Address, Country, WWW.

I. INTRODUCTION

With the dramatically quick and explosive growth of information available over the Internet, World Wide Web has become a powerful platform to store, broadcast and retrieve information as well as mine useful knowledge. Due to the properties of the huge, diverse, dynamic and unstructured nature of Web data, Web data research has encountered a lot of challenges, such as scalability, multimedia and temporal issues etc. As a result, Web users are always drowning in an "ocean" of information and facing the problem of information overload when interacting with the web. Typically, the following problems are often mentioned in Web related research and applications:

Finding relevant information: To find specific information on the web, users often either browse Web documents directly or use a search engine as a search assistant. When a user utilizes a search engine to locate information, he or she often enters one or several keywords as a query, then the search engine returns a list of ranked pages based on the relevance to the query. However, there are usually two major concerns associated with the query-based Web search [1]. The first problem is low precision, which is caused by a lot of irrelevant pages returned by the search engine. The second problem is low recall, which is due to the lack of capability of indexing all Web pages available on the Internet. This causes the difficulty in locating the un-indexed information that is actually relevant. How to find more relevant pages to the query, thus, is becoming a popular topic in Web data management in last decade.

Finding needed information: Most search engines perform in a query-triggered way that is mainly on a basis of one keyword or several keywords entered. Sometimes the results returned by the search engine don't exactly match what a user really needs due to the fact of the existence of the homology.

Learning useful knowledge: With traditional Web search service, query results relevant to query input are returned to Web users in a ranked list of pages. In some cases, we are interested in not only browsing the returned collection of Web pages, but also extracting potentially useful knowledge out of them.

Recommendation/personalization of information: While a user interacts with the web, there is a wide variety of user's navigational preference, which results in needing different contents and presentations of information. To improve the Internet service quality and increase the user click rate on a specific website, thus, it is necessary for a Web developer or designer to know what the user really wants to do, predict which pages the user is potentially interested in, and present the customized Web pages to the user by learning user navigational pattern knowledge.

The above problems place the existing search engines and other Web applications under significant stress. A variety of efforts have been contributed to deal with these difficulties by developing advanced computational intelligent techniques or algorithms from different research domains, such as database, data mining, machine learning, information retrieval and knowledge management etc. Therefore, the emerging of the Web has put forward a great number of challenges to Web researchers and engineers for web-based data management and Web application development.

II. WEB MINING TECHNIQUES

Nowadays the Internet has been well known as a big data repository consisting of a variety of data types as well as a large amount of unseen informative knowledge, which can be discovered via a wide range of data mining or web mining techniques [2]. All these kinds of techniques are based on intelligent computing approaches, or so-called computational intelligence, which are widely used in the research of database, data mining and information retrieval and so on. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The most popular approach in the direction of discovering such knowledge through log files is that of Web mining. The aim of discovering such knowledge through log data is to obtain information about the navigational behavior of the users. This can be used for advertising purposes, for creating dynamic user profiles etc.

Web mining [3,4] is to explore interesting information and potential patterns from the contents of Web page, the information of accessing the Web, page linkages and resources of E-commerce by using techniques of data mining, which can help people extract knowledge, improve Web sites design, and develop e-commerce better.

The usage of the data mining [5] process to these dissimilar data sets is based on the three different research directions in the area of web mining: web content mining, web structure mining and web usage mining.

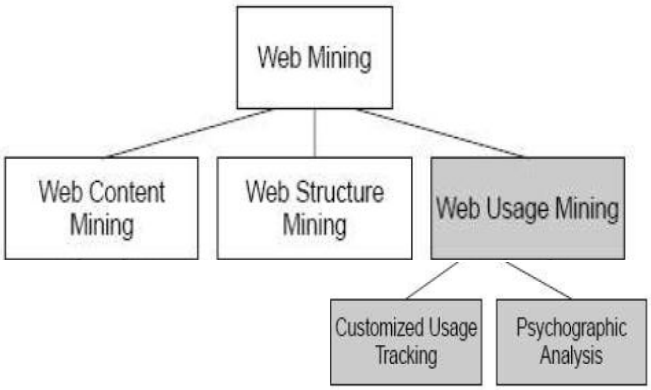


Figure1-Taxonomy of Web Mining

WEB CONTENT MINING-Web content mining [6, 7] is the process of extracting knowledge from the content of a number of web documents. Web structure mining is the process of inferring knowledge from the organization and links on the web, while web usage mining is automatic discovery of user access patterns from web servers.

WEB STRUCTURE MINING- Structure Mining [5, 6] aims at finding the underlying topology of the interconnections between web objects. Web structure mining is concerned with discovering the model underlying the link structures of the web. it is used to study the topology of the hyperlinks with or without the description of the links. This model can be used to categorize web pages and is useful to generate information such the similarity and relationship between different web sites.

WEB USAGE MINING- Usage Mining [8] is the application of data mining techniques to discover usage patterns from web data. Data is usually collected from user's interaction with the web, e.g. web/proxy server logs, user queries, registration data. Usage mining tools discover and predict user behavior, in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service [9].

Web usage mining [10] deals with studying the data generated by the web surfer's sessions or behaviors. We know that the Web content and structure Mining utilize the real or primary data on the web. On the contrary, Web usage mining mines the secondary data derived from the interactions of the users with the web. The secondary data includes the data from the web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions cookies user queries, bookmark data, mouse clicks and scrolls and any other data which are the result of these interactions.

This data can be accumulated by the web server. Analysis of the web access logs of different web sites can facilitates an understanding of the user behaviour and the web structure, thereby improving the design of this colossal collection of information. There are two main approaches in web usage mining driven by the applications of the discoveries.

PSYCHOGRAPHIC ANALYSIS

CUSTOMIZED USAGE TRACKING

It is important to note that the success of such applications depend on what and how much valid and reliable knowledge one can discover from the large, raw log data. For effective web usage mining, an important cleaning and data transformation step may be needed before analysis.

Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior and the web structure, thereby improving the design of this colossal collection of resources. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking.

The general access pattern tracking analyzes the web logs to understand access patterns and trends. These analyses can shed light on better structure and grouping of resource providers. Many web analysis tools existed but they are limited and usually unsatisfactory.

Customized usage tracking analyzes individual trends. Its purpose is to customize web sites to users. The information displayed the depth of the site structure and the format of the resources can all be dynamically customized for each user over time based on their access patterns. While it is encouraging and exciting to see the various potential applications of web log file analysis, it is important to know that the success of such applications depends on what and how much valid and reliable knowledge one can discover from the large raw log data. Current web servers store limited information about the accesses. Some scripts custom-tailored for some sites may store additional information. However, for an effective web usage mining, an important cleaning and data transformation step before analysis may be needed.

III. PROPOSED WORK

''Frequency'' and ''Duration'' are the only two factors in the weight measure for capturing the interest degree of a web page to a user. Due to they are considered two strong indicators for capturing the interest degree of a web page to a user. Some other implicit factors may also indicate users' interests and preferences in web mining. We proposed a third factor "Country" to indicate users' interests and preferences.

Every user belongs from a particular Country and every Country has some particular religion, traditions, thinking, faith, market, need of population, customer requirement for example Indian user believes mostly on astrology so content of web page can be changed according to users' Country.

IV. METHODOLOGY USED

Our work will help to personalize the website according to users' "Country". We know that every internet user use one unique IP Address and that IP Address is provided by the ISP(Internet Service Provider).every region has some particular ISP[11] for example for India region BSNL is the major ISP. Through IP Address we can find out the Country of the user and then we can change the content of web pages according to Country to predict the users' future request.

To find out the Country or location of the user we have only one thing that can be uniquely identify the user location that is IP address of the device from which he\she is accessing the internet. Through IP Address we can find out the accurate Country of the user and then we can change the content of web pages according to Country for improving the users' future request. Now our priority to find out the end user Country from where he\she access the WWW. For this purpose we have to concentrate on IP Address identification. After

that we will take a server database (Web log data) for experiment and analyze that data and find a general user access pattern or preferences according to "Country" and then personalize a website according to a user need. After that to increase the accuracy of the results we apply the A priori Algorithm [6, 12] and find out the frequent item set to calculate the Support and Confidence for particular URL or web-pages. From these results we can find out the most visited pages in a particular web-site and we can observe the user interest's and can customize our web-site according to the user from a particular Country.

V. FRAMEWORK FOR IMPLEMENTATON

The frame work for Web Data Mining using site's Web server logs as data source is shown in figure 2-

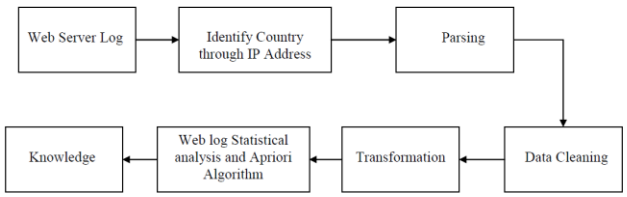


Figure 2- Frame work for implementation

Explanation of this Framework for Proposed Work is as follows:-

- We will collect or arrange a Web server access log (Web Log) of any Web-site.
- In the second step we will identify the User's Country from the IP address means we will identify that user belong to which region?
- In the third step Parsing [13, 15] will be performed on that database means apply the method of converting the unstructured data (raw data) generated from web server log into structured data (HTML table or Data base table).
- The primary objective of Web mining is to discover interesting patterns in accesses to various Web pages within the Web space associated with a particular server [14]. The server logs contain the entries that are redundant or irrelevant for those web mining tasks. For example, all the image files entries are relevant or redundant. As a URL with several image files is selected, the images are transferred to the client machine and these files are recorded in the log file as independent entries. For our purposes, image files without a hyper link can be ignored because they are not used to visit another pages.

An image file with a hyper link is redundant since as a client selects the hyper link, the destination URL will be recorded in the log. A similar situation exits with regards to image map files. Unless the task is to determine which display is better to attract the clients to visit certain URLs, we can remove all these redundant log eateries for the Web mining tasks. We call this process Data cleaning.

Data cleaning [13, 15] is performed by checking the suffix of the URL name. All the URL entries with filename suffixes such as jpeg, jpg and map are removed from the log. Data cleaning process also involves the selection of a subset of the fields that are relevant for the data mining tasks. The fields of interest are IP address, user id, access time and the URL .After the log entries are cleaned; the log data is converted into a form suitable for a specific Web mining task. So in this step of Data Cleaning we remove the url which contain the images as a destination.

- In the next step, after the log entries are cleaned, the log data is converted into a form suitable for a specific Web Mining Tasks or the data are transformed or consolidate into forms appropriate for Web mining [13].

  This transformation is accomplished according to the transaction model for that particular task (e.g., association rule) to get the required result and analysis.
- Log analysis [16] is regarded as one of method used in the Web Mining Process. The purpose of Web Mining is to apply statistical and data mining techniques to the pre-processed Web log data, in order to discover useful pattern. The most common and simple method that can apply to such data is statistical analysis.

  In this step we will apply Web log Statistical analysis on that database to find out Frequency of Visitors from different Countries, Page accessed Frequency from a particular Country, Number of Visitors per Hour, Most visited pages in a web-site.
- In the second last step we will apply A-priori Algorithm to the database to compute frequent item sets and then determine the Support and Confidence for different url of a web-site.
- Knowledge Discovery [17] is the final step in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the web analysis result. They can be displayed as raw data, tables, graphs, bar-chart, rules and pie-charts. This process is tried to make the analysis results easier to be used and more understandable.

Steps to accomplished our Implementation

i. Collection of Web Server Log

We took raw data of Nasa's Web-site of July month, 1995 from web site www.web-caching.com.

ii. Identify Geographical Location through IP Address

User can be identified through cookies, logins, or IP/agent/path analysis. To identify the user a client side tracking mechanism is used, only the IP address, agent and server side click stream are available to identify users.

iii. Log Parsing

Log Parsing means apply the method of converting the unstructured data (raw data) generated from web server log

into structured data (HTML table or Data base table).We convert this raw web log data into structured data(in to table) in My-SQL server through java programming.

iv. Data Cleaning

In this step we remove the url which contain the images as a destination. To do this we run sql queries through java coding.

v. Data Transformation

After the log entries are arranged into the MS-Access database form, the log data is converted into a form suitable for a specific Web analysis Task [16]. This transformation is accomplished according to the transaction model for that particular task for ex convert the Link field into Link-id and add a Country field into database according to IP Address for running a SQL queries to get the required result and analysis.

vi. Web Log Analysis

Log analysis is regarded as one of method used in the Web Mining Process. The purpose of Web Mining is to apply statistical and data mining techniques to the pre-processed Web log data, in order to discover useful pattern. The most common and simple method that can apply to such data is statistical analysis.For analysis of this database we used MATLAB 7.0.we used one of the database tool which is "Visual Query Builder". From this tool we made lot of different SQL queries on this database for recognization of user's interest and stored these results in different MATLAB variables. After that we made a Bar-Charts, Graphs and Pie-Chart for different parameters like Country, Hour, Links accessed, No. of Visitors and frequency of different parameters to extract the following information-

- Prediction of the user's behavior within the site.
- General behavior patterns across all users.
- Adjustment of the Web-site to the interest of its users.
- Prediction of the user's behavior according to Geographical location.
- Prediction of the user's behavior according to most active country.

vii. Knowledge Discovery

Knowledge Discovery [17, 18] is the final step in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the web analysis result. They can be displayed as raw data, tables, graphs, bar-chart, rules and pie-charts. This process is tried to make the analysis results easier to be used and more understandable. We can personalize Our Website's according to user's interest.

viii. Results Analysis

All the experiments were performed on 1.8 GHz Pentium PC with 512 MB RAM, running Windows-XP. This work is done by one of the Database tool "Visual Query Builder" of MATLAB 7.0.

Result of Access Statistics

By Most Popular url or link or Pages

To know the most popular url, pages or link among the Visitors. We will run a SQL query on the table Nasa. After running this query result is stored in MATLAB variable and draws a graph between Links Accessed and No of Visitors by the help of Visual Query Builder. The result of this query is shown below in form of graph. From with this result we can say that link L4 is most popular link.

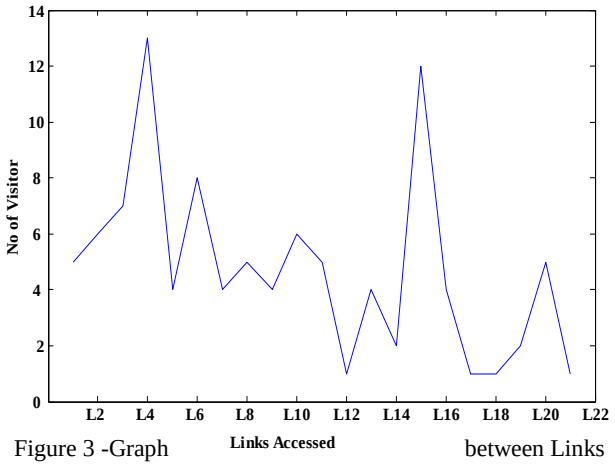


Figure 3 -Graph between Links Accessed Vs No. of Visitors

If we want to know the accessed frequency of links by USA Visitors then we have to run the query. Result of the query is stored and drowns a Bar-Chart between Links Accessed vs. Frequency by assist of Database tool of MATLAB "Visual Query Builder".

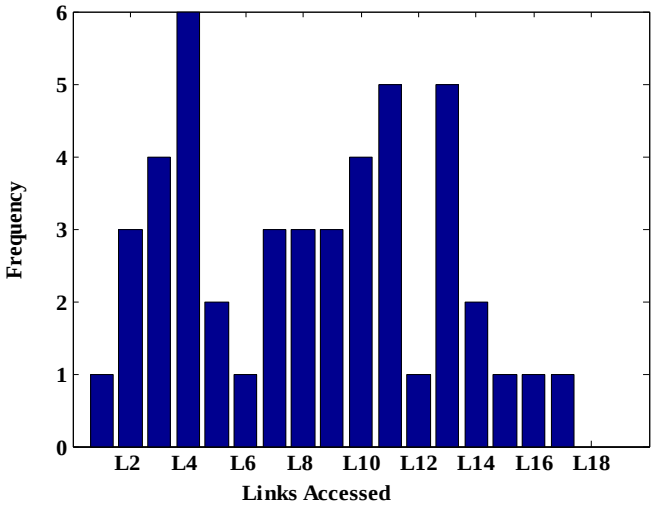


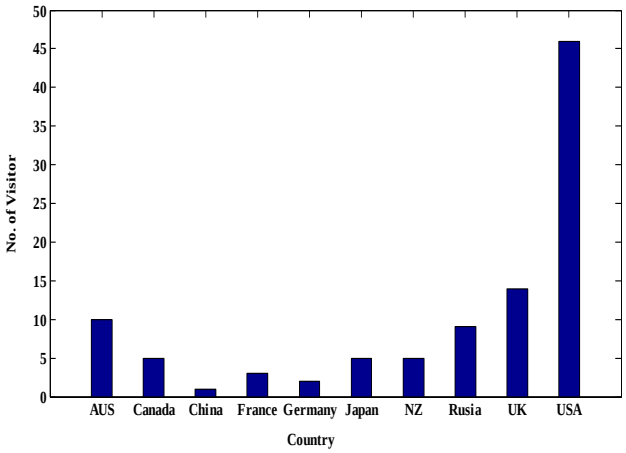
Figure 4 -Bar Chart for Links Accessed Frequency by USA Visitors

Result of Countries Statistics

By Most Active Country

To draw Bar-chart between Countries vs. No of Visitors from these countries we will run a SQL query and apply same procedure to Visual Query Builder to draw Bar-chart. From this result we can see that Highest No of visitors from USA country and after that UK visitors access this site most among all countries which accesses this site.

Figure 5-Bar Chart for Most Active Country



To draw Pie-Chart of countries to know which country is most active among all these countries we will run same SQL query. A result is stored in MATLAB variable and then goes display option and choose chart option and after that select a pie chart among all options. From Pie-Chart we conclude that USA is most active country among all countries which accesses this site.

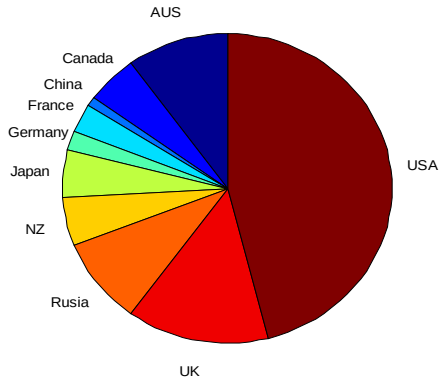


Figure 6-Pie-Chart for most active country

Future Enhancement

To extract user navigation behavior ‘‘Frequency’’ and ‘‘Duration’’ were the only two factors in the weight measure due to they are considered two strong indicators for capturing the interest degree of a web page to a user. Now we have proposed third factor “Geographical location” in this Paper for capturing the interest degree of a web page to a user. Some other implicit factors, for example

- ❖ The Time zones of accessed web page may also indicate users’ interest and preferences and we can add this factor also to know the interest degree of a user in a particular time zone.
- ❖ The sequence of accessed web pages may also indicate users’ interests and preferences. We are interested to add this factor also into the weight measure of the interest degree of a web page to a user.

CONCLUSION

The web is a very essential means to carry out business and commerce so the design of web pages is highly essential for the system managers and web creators. These characteristics have huge impact on the number of users who access the page. Therefore the web analyzer has to examine with the data of server log file for identifying the navigation pattern.

The implemented framework is efficiently discovering the knowledge from web server log data using web log analysis .

By applying Web log Statistical analysis, we will get web analysis result in the form of Bar-Charts, Pie-Chart, Graphs and association rules on the basis of different parameters like Country, Links accessed, No. of Visitors and frequency of different parameters to extract the following information-

i. Prediction of the user’s behavior within the site.

ii. General behavior patterns across all users.

iii. Adjustment of the Web-site to the interest of its users.

iv. Prediction of the user’s behavior according to Country.

v. Prediction of the user’s behavior according to most active country.

Our paper will provide a frame work to predictions to web site designer based on discovered rules from web server log on the basis of Country. Considering these rules web site designer can make appropriate adjustment into the index page and URL links effective and easy navigation through web sites.

The Generated association rules from web server log may be used to predict the next likely HTTP request from end user.

The evaluation of pattern (according to country) can be carried out in an integrated manner for problems like link prediction. It is obvious that enhanced pattern discovery provides highly accurate guessing of a Web user’s future visit if the user’s pattern can be exactly determined.

REFERENCES

1. Kosala, R. and H. Blockeel, *Web Mining Research: A Survey*. SIGKDD Explorations, 2000. 2(1): p. 1-15.
2. Renáta Iváncsy, István Vajk, "Frequent Pattern Mining in Web Log Data", Acta Polytechnica Hungarica, Vol. 3, No. 1, 2006, Pp 77-90.
3. Haibin Liu, Vlado Kesˇelj . “Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users’ future requests”. Data & Knowledge Engineering 61 (2007) 305–330.
4. Federico Michele Facca, Pier Luca Lanzi ”Mining interesting knowledge from web logs: a survey”. Data & Knowledge Engineering 53 (2005) 225–251.

5. Han, J. and M. Kamber, Data Mining: Concepts and Techniques. 2007: Morgan Kaufmann.
6. Arun K Pujari, Data Mining Techniques, Chapter 8, Edition-2007.
7. Liu, B. and K.C.-C. Chang, "*Special Issue on Web Content Mining"*. ACM SIGKDD Explorations, 6(2): Pp. 1-4, 2004.
8. Mobasher, B., Web Usage Mining and Personalization, in Practical Handbook of Internet Computing, M.P. Singh, Editor. 2005, CRC Press. p. 15.1-37.
9. Eirinaki M., Vazirgiannis M. (2003). Web mining for web personalization. ACM Transactions on Internet Technology (TOIT), 3(1), 1-27.
10. Srivastava, J., et al., Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM SIKDD Explorations, 1 (2): p. 12-23. Internet Computing, M.P. Singh, Editor. 2005, CRC Press. p. 15.1-37.
11. http://www.apnic.net/
12. R. Agrawal,R. Shrikanth, "Fast Algorithm for mining Association Rule" Proc. of VLDB Conference, pp.587-559, Santigo, Chile, 1995.
13. Tanasa, D.; Trousse, B.; "Data preprocessing for WUM", IEEE Potentials, Vol. 23, No. 3, Pp. 22 – 25, 2004.
14. Srivastava, J.;"Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data". ACM SIKDD Explorations, 1(2): Pp. 12-23, 2000.
15. R. Cooley, B. Mobasher and J. Srivastava Data preparation for mining World Wide Web browsing patters. Knowledge and Information Systems, 1(1), 1999.
16. Nikos Koutsoupias, " Exploring Web Access Logs with Correspondence Analysis", 2nd Hellenic Conf. on AI, SETN-2002, Thessaloniki, Greece, Proceedings, Companion Volume, Pp. 229-236, 11-12 April 2002. COOLEY, R., TAN, P-N., AND SRIVASTAVA, J. 1999b. "WebSIFT: The web site information filter system. In Proceedings of the Web Usage Analysis and User Profiling" Workshop (WEBKDD'99), Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Boston, August).
17. COOLEY, R., TAN, P-N., AND SRIVASTAVA, J. 1999b. "WebSIFT: The web site information filter system. In Proceedings of the Web Usage Analysis and User Profiling" Workshop (WEBKDD'99), Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Boston, August).