IJCSE International Journal of Computer Science and Engineering Open Access
Review Paper Volume-2, Issue-5 E-ISSN: 2347-2693

A Review of Optimisation of Search Engine using Sequential Pattern Mining Technique

Vandana Dhull[1*] and Shipra Khurana[2]
[1*,2]*Department of Computer Science & Engineering, DVIET, Karnal, India*
www.ijcseonline.org
Received: 11/04/2014 Revised: 10 /05/2014 Accepted: 20/05/ 2014 Published: 31 /05/2014

***Abstract*—**with the large increase in the amount of information available online, rich web data can be obtained on the internet, such as over one trillion. Web mining techniques has emerged as an important research area to help web users find their information need. Web user express their information need as queries, and expect to obtain the needed information from the web data through web mining technique. Nowadays, providing an amount of relevant web pages based on users query words is a not a big problem in search engines. Instead, the problem is that a search engine returns too many web pages, and users have to spend much time on finding their desired information from this long search result list, named as Information Overloaded Problem. Finally, search result list is re-ranked by modifying the page rank algorithm using the weights assigned to sequential patterns resulting in reduction of users navigation time within the search result.

Keywords- *Information; Web Mining; Web Pages; Search Engine; Patterns; Navigation Time; Page Rank Algorithm*

I. INTRODUCTION

The information space known as Web is a collection of resources (Web Patterns) residing on the Internet, that can be accessed using Hyper Text Transfer Protocol (HTTP) and protocols that derive from it. Resource can be anything that has identity. When a resource is accessed by a client at a specific time and space, then it is called as resource manifestation. The general definition for client according to [2] is "an application capable of accessing Web resources by issuing requests and render responses containing Web resource manifestations". Uniform Resource Identifier (URI) defines a "compact string of characters for identifying an abstract or physical resource" [2].

When talking about Web resources, concept of a subset of the URI space, called URLis also important. The term Uniform Resource Locator refers to the "subset of URI that identify resources via representation of their primary access mechanism rather than identifying the resource by name of that resource." For example: http://www.yahoo.com and http://www.altavista .com are examples of URLs which identify the main page of two most famous Internet Web site. With the advancement in information technologies, the Web has become a huge information repository that covers almost all the topics in which a human user could be interested. In spite of the recent advances in the Web Search engine technologies; there are still many situations in which the user is presented with non-relevant search results. One of the major reasons for this problem is that most Web search engine users are not well trained in organizing and formulating their input queries, on which the search engine relies to find the desired search results. Moreover, Web search engines often have difficulties in forming a concise and precise representation of the user's information need. Although many search engines provide a user friendly ranked list in response to user queries, still there remains a challenge in ranking the document's relevance based on user queries. Web mining is a technique that can help to tackle this challenge.

Corresponding Author: *Vandana Dhull*

II. WEB MINING AND ITS TAXONOMY

Web mining is the application of data mining technique to extract knowledge from Web data, where at least one of content, structure (hyperlink) or usage (Web log) data is used in the mining process (with or without other types of Web data). The process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video or structured records such as lists and tables.

- Web structure Mining (WSM)

Web structure mining is the process of discovering structural information from the Web. The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages.

- Web Usage Mining (WUM)

It is the application of data mining technique to discover interesting usage patterns from Web data. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. . Colley et.al. gave definition of web usage mining[4] " WUM is the "*automatic discovery of user access patterns from Web servers".*

III. ARCHITECTURE OF WEB USAGE MINING

- *Log Information:*

A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site, the types of information the server preserves include the user's domain, sub domain, hostname, the time of the request, and any errors returned by the server.

Following list shows the information stored in a Web log:

- Username and password if the server requires user's authentication.
- Bytes: The content-length of the document transferred.
- Entering and exiting date and time.

- Remote IP address or domain name: An IP address is a 32-bit host address defined by the Internet Protocol and a domain name is used to determine a unique Internet address for any host on the Internet.
- Remote log and agent log.
- Remote URL
- "Request" The request line exactly as it came from the client.
- Requested URL.
- Status: The HTTP status code returned to the client, e.g. 200 is "ok" and 404 are "not found".

The CGI environment variables [11] supply values for many of the above items.
Web server log files were used initially by the webmasters and system administrators for the purpose of record and trace the visitors' on –line behaviors. Web log file is one way to collect Web traffic data. The other way is to "sniff" TCP/IP packets as they cross the network, and to "plug in" to each Web server. After the Web traffic data is obtained, it may be combined with other relational databases, over which the data mining techniques are implemented. Through some data mining technique such as association rules, path analysis, sequential analysis, clustering and classification, visitors' behavior patterns are found and interpreted.

- *Query Log*

The query log is a text file consisting of a series of requests. A request may consist of a new query or a new result screen for a previously submitted query. Each request includes the following fields.

- AnonID: An anonymous user ID number, usually corresponding to a real search engine user.
- Query: the query issued by the user.
- Query Time**:** the time at which the query was submitted to the search engine by the user for fulfilling his particular information needs.
- Item Rank**:** If the user clicked on a search result, the rank of the item on which they clicked is listed.
- Click URL**:** if the user clicked on a search result, the domain portion of the URL in the clicked result is listed.

- *Web Log Information*

A Web log file records activity information when a Web user submits a request to a Web server. A log file can be located in three different places:

1) Server-side logs
2) Proxy-side logs
3) Client -side logs

The above mentioned each file is illustrated below with major drawbacks of each:

1) Server-side logs**: -** These logs generally supply the most complete and accurate usage data. But their two drawbacks are:
   - These logs contain sensitive, personal information, therefore the server owners usually keep them closed.
   - The logs do not record cached pages visited. The cached pages are called from local storage of browsers or proxy servers, not from Web servers.

2) Proxy-side logs**:** A proxy server takes the HTTP requests from users and passes them to a Web server; the proxy server than returns to users the results passed to them by the Web server. The two disadvantages are:
   - Proxy server construction is a difficult task. Advanced network programming, such as TCP/IP is required for this construction.
   - The request interception is limited, rather than covering most requests.

The proxy logger implementation in Web Quilt [8], a Web logging system, can be used to solve these two problems, but the system performance declines because each page request needs to be processed by the proxy simulator.

3) Client-side logs**:** Participants remotely test a Web site by downloading special software that records Web usage or by modifying the source code of an existing browser. HTTP cookies could also be used for this purpose. These are pieces of information generated by a Web server and stored in the users' computers, ready for future access. The drawbacks of this approach are:

- The design team must deploy the special software and have the end-users install it.
- This technique makes it hard to achieve compatibility with a range of operating systems and Web browsers.

Content preprocessing consists of converting the text, image, scripts, and other files such as multimedia into forms that are useful for the WUM process. Content preprocessing performing classification or clustering. Contents of a site can be used to filter the input, output from the pattern discovery algorithms.

Page views can be intended to convey information, gather information from the user, and allow navigation. Page views can also be classified according to their intended use [5, 12].The intended use of a page view can also filter the sessions before or after pattern discovery.

In order to run content mining algorithms on page views, the information must first be converted into a quantifiable format. Some version of the vector space model [13] is typically used to accomplish this.

Structure Preprocessing
The structure of a site is created by the hypertext links between page views. The structure can be obtained and pre-processed in the same manner as the content of a site. A different site structure may be constructed for each server session.

Sequential pattern discovery
Sequential patterns discovery is to find the inter-transaction patterns such that the presence of a set of items is followed

by another item in the time-stamp ordered transaction set. Web log files can record a set of transactions in time sequence. If the web-based companies can discover the sequential patterns of the visitors, the companies can predict user's visit patterns and target market on a group of users. The sequential patterns can be discovered as the following form:
50% of client who bought items in/pcworld/computers/, also placed an order online in/pcworld/accessories/within 15days

Clustering
Clustering identifies visitors who share common characteristics. Customer profiles often need to be obtained from an online survey form when the transaction occurs. For example, one may be asked to answer the questions like age, gender, email account, mailing address, hobbies etc. Those data will be stored in the company's customer profile database, and will be used for future data mining purpose. An example of clustering could be:
50% of clients who applied discover platinum card in /discovercard/customer Service/newcard, were in the 25-30 age group, with annual income between $40,000-50,000.

Clustering of client information can be used on the development and execution of future marketing strategies, online and off-line, such as automated mailing campaign.

Decision Trees
A decision tree is essentially a flow chart of questions or data points that ultimately leads to a decision. Decision trees systems are incorporated in product-selection systems offered by many vendors.

Pattern Analysis
Pattern analysis is the last step in the overall WUM process. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations.

Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data. The end products of such analysis might include:

1. The frequency of visits per document.
2. Most recent visit per document
3. Who is visiting which document?
4. Frequency of use of each hyperlink, and
5. Most recent use of each hyperlink.

The common techniques used for pattern analysis are visualization techniques, OLAP techniques, Data & Knowledge Querying and Usability Analysis.

IV . QUERY CLUSTERING, SEQUENTIAL PATTER MINING & PAGE RANKING

Query clustering: Query clustering is one of the techniques used in query mining, and it groups similar queries automatically without using predetermined class descriptions. It allows users to utilize other users' search experience or domain knowledge by analyzing the information stored in query logs, and then grouping and extracting useful related information on a given query.
Although the need for query clustering (33, 46) is relatively new, there have been extensive studies on document clustering, which is similar to query clustering.
Approaches used for Query Clustering are as follows:

Keyword based approach
In these approaches, a document is represented as a vector in a vector space formed by all the keywords [13].Keyword based approach, Clustering documents using the keywords they contain.
Keyword-based document clustering has provided interesting results. One contributing factor is the large number of keywords contained in documents. Even if some of the keywords of two similar documents are different, there are still many others that can make the document similar in the similarity calculation.

Limitations:
The queries submitted to the search engines, typically are very short, in many cases it is hard to deduce the semantics from the queries themselves. Therefore, keywords alone do not provide a reliable basis for clustering queries effectively. In addition, words such as 'where" and "who" are treated as stop words in traditional IR methods. For example, with a "who" question, the user intends to find information about a person. So even if a keyword-based approach is used in query clustering, it should be modified from that used in traditional clustering.

Hyperlink based clustering
Because of the limitations of keyword based approaches, hyperlinks [5] between documents are required. The hypothesis is that hyperlinks connect similar documents. This idea has been used in some early studies in IR [22]. More recent examples are Google (http://www.google.com) and the authority/hub calculation of Kleinberg.

Limitations
In the hyperlink approaches, document space and query space are still separated.

Cross-reference between Queries and Documents
By cross-reference, it means any relationship created between a query and a document. The intuition of using cross-reference is that similarity between documents can be transferred to queries through these references, and vice-versa.
Relevance feedback in IR is a typical exploitation of cross-reference. It is typically used to reformulate the user's query [13].

In the Web environment, the choice of a particular document from the result list by a user is another kind of cross-reference between queries and documents. Although it is not as accurate as explicit relevance judgment in traditional IR, the user's choice does suggest a certain degree of "relevance" of that document to his information need. In fact, users usually do not make the choice randomly. The choice is based on the information provided by the search engine.

A similar idea can be used for query clustering – if a set of queries often lead to the same or similar document clicks, and then these queries are similar, to some extent. These ideas have been used in some work in the area of IR.

Sequential Pattern Making

Sequential Pattern Making is one of the primary techniques of Data Mining. It is the mining of frequently occurring ordered events or subsequences as patterns. For retail data, sequential patterns are useful for shelf placements and promotions. The industry, as well as telecommunications and other businesses, may also use sequential patterns for targeted marketing, customer retention and many other tasks. Other areas in which sequential patterns can be applied include Web access pattern analysis, weather prediction, production processes and network intrusion detection.

The Problem of Sequential Pattern Mining

The sequential pattern mining problem was first introduced by Agarwal and Srikant in 1995 [23] based on their study of customer purchase sequences, as follows:

"Given a set of sequences, where each sequence consists of a list of events (or elements) and each event consists of items, and given a user-specified minimum, support threshold of min_sup, sequential pattern mining finds all frequent subsequences, that is, the subsequences whose occurrence frequency in the set of sequences is no less than min_sup".

An example of a sequence data is given in Table 2.1 Here, the input data is a set of sequences, called data-sequences. Each data-sequence is a list of transactions, where each transaction is a set of literals, called items. Typically, there is a transaction-time associated with each transaction. A sequential pattern also consists of a list of set of items. The problem is to find all sequential patterns with a user-specified minimum support, where the support of a sequential pattern is the percentage of data-sequences that contain the pattern.

Table 2.1: A real-time example of a sequence data

Sequence Database	Sequence	Element (Transaction)	Event (item)
Customer	Purchase history of a given document	A set of items bought by a customer at time t	Books, diary products, CDs, etc.
Web Data	Browsing activity of a particular Web visitor	A collection of files viewed by a Web visitor after a single mouse click	Home pages, index page, contact info. Etc.

However, the above given problem definition has the following limitations:

Absence of time constraints: User often wants to specify maximum and/or minimum time gaps between adjacent elements of the sequential pattern.

Rigid definition of a transaction: For many applications, it does not matter if items in an element of a sequential pattern were present in two different transactions, as long as the transaction-times of those transactions are within some small time window. Empire' and 'Ringworld Engineers' ".

Absence of taxonomies: Many datasets have a user-defined taxonomy (is-a hierarchy) over the items in the data, and users want to find patterns that include items across different levels of the taxonomy.

V. REQUIREMENT FOR OPTIMISATION

The objectives of the proposed work are given below

- *Query-based approach* is used which is dependent on user's query words and will provide search results in a relevant order according to user's intent.
- *To overcome the shortcoming of traditional ranking methods* as it is based on content oriented and link oriented approaches.
- *To reduce the time users spend for seeking out their required information from search result list* as the relevant pages may have low ranking in the search result list.
- *To generate search results more relevant to a user's intent* as out of thousands of search results, only few are read by the user.

To effectively refine the ranking of search results for any given querry by constructing the query context from search query logs so that the relevant pages will get higher ranking as compared to irrelevant results.

VI OPTIMISATION OF SEARCH ENGINE USING SEQUENTIAL PATTERN MINING TECHNIQUE

A novel rank updation method "Rank Boosting" is proposed that uses the logs of search engines to boost their retrieval quality. The relevance of Web pages is estimated using the historical preferences of user that appear in the logs. The purpose is to cluster queries according to the similarity based on URLs clicked in their answers. After that specific ranking models are trained for each query cluster using the frequent pattern generated from pattern mining algorithm. The proposed approach is tested using query log data from a search engine. It turns out that the proposed topic-dependent models can significantly improve the search results. The flowchart for the proposed work is represented in Figure 1.1.

According to the flowchart given in Figure 1.1, firstly, there is a need to extract the session from the query log. Session consists of the queries issued by the user and the corresponding document clicks. Since this large amount of data may consists of noise and missing values, thus preprocessing needs to be done. After this, the similarity needs to be determined among the queries based on user's queries and corresponding document clicks and get clustered in a database using Query Clustering. Now, a sequential pattern mining algorithm i.e. GSP is implemented on the data to generate frequent pattern and finally the rank of each URL get updated with the help of Weight Calculator by calculating the weighted level of URL in a generated pattern hierarchy and Page Rank algorithm. A new approach is proposed for query clustering using logs which could enhance the quality of search results. In particular, the cross-references between the users' queries and the documents that the users have chosen to read is considered. The hypothesis is that there is a strong relationship between the queries and the selected documents (or clicked documents).
In most search engines, users' searching behaviors including search results, Web pages' information, and browsing activities are recorded in the query logs. These query logs contain rich information, which can be used to analyze users' behaviors and improve the search performance.

Based on this information, initially the query sessions are constructed which consists of a query, and the corresponding clicked Web pages from query logs. For each set, the original ranking order with each clicked Web page is also extracted from the source search engine's query logs. In addition, it is assumed that a Web page clicked by many users with users with the same query word might have a higher-ranking score, named as users' feedback ranking score. It is a very intuitive concept that higher clicked times means more relevant because most user recommend those Web pages with requirement. Moreover, the Web pages shown at the top of the search results list form the source search engine are assumed to have higher-ranking score, named as content-oriented ranking score. Therefore, if a Web page can satisfy these two ranking concepts, it will get a higher score for ranking. Then a similarity function (based on user's content and feedback) has been applied on the queries to perform query clustering.

The main aspect of this dissertation is not only to perform query clustering based on user's query keywords and feedback but also to generate frequent patterns of web pages requested by a user for a particular query with the help of Generalized Sequential Pattern (GSP) algorithm. The final approach is to re-rank the search result list by modifying the page rank algorithm using the weighted levels of sequential patterns so that the time users spend for seeking out their required information from search result list can be reduced and the more relevant Web pages can be presented.
The proposed architecture consists of the following functional components :

1. Query Similarity
2. Query Clustering
3. Pattern Generator
4. Rank Boosting

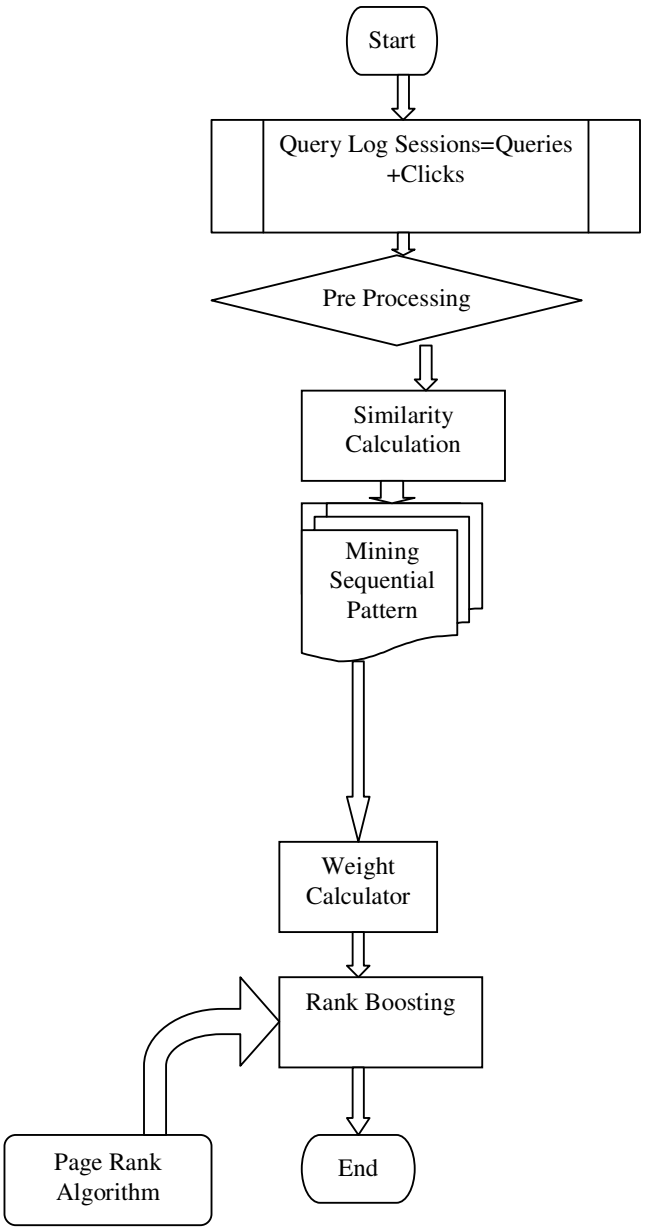

Figure 1.1 Flowchart of Search Result Optimization

The notion of the algorithm is based on the simple perspective as given below:

1) Initially, a cluster Id 0 is assigned to each query which represents that a particular query is not assigned to any cluster.
2) Now, each query is checked for its cluster id. If its value is 0, then the similarity between it and the next query (which has not been assigned to any cluster yet) is calculated using the formula given in (4.3).
3) If the resulting value satisfies threshold value i.e. min_sup the given query is assigned to a particular cluster along with its Cluster Id.
4) Steps 2 and 3 are repeated for each query
5) After creating the clusters, each query is sorted according to its cluster Id and finally the whole result gets stored in Query Cluster database.

To implement the proposed architecture, a sample Query Log is considered, from whole user query sessions are extracted. Because the actual number of queries is too big to conduct detailed evaluation, only 14 query sessions are chosen from them. The following functions are tested on the 14 query sessions:

- Keyword similarity ($Sim_{keyword}$).
- Similarity using documents clicks (Sim_{click})
- Similarity using both keyword and document clicks ($Sim_{combined}$)
- Query clustering
- GSP algorithm
- Rank boosting
-

In the third function, both a and b are set to 0.5 and min_sup = 2

The various steps performed on the sample logs are given below:

Step 1: First of all, query similarity is determined with the help of Query_Sim algorithm given in Figure 5.3.

Step 2: by taking the URLs from a Query Cluster Database for a particular cluster id, a pattern will be generated with the help of a GSP algorithm (Refer Figure 5.5). first of all, two columns (Cluster Id, URL) need to be retrieved from the Table 5.3 for a cluster Id 1. to simplify the calculation. URLs are assigned to different variables.

D= www.cardekhop.com and so on. And then the refined data in Table 5.4 need to be analyzed to generate frequent patterns with the help of GSP algorithm.

Step 3: From the above result, a weight of each URL can be determined. The generated pattern can be represented graphically as can be seen from figure 5.8

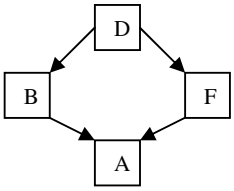

Figure 1.2 Graphical representation of generated pattern

Since in Figure 1.2, D is at first level, hence it's a weight according to $W_1 = 1/L$, is 1/1. In the similar way, the weight of B,F and A is ½, ½, 1/3

Now, for modifying the rank, the proposed Rank Boosting algorithm. Page Rank of URLs can be taken from Query log. To illustrate here, let us say the page rank of URLs are:

PR (A) =9, PR(B)=8, PR(D)=6, PR(F)=4, then according to

New Rank (D) = PR(D) *W_t(D) = 6 * 1/1 = 6

New Rank (B) = PR(B) *W_t(B) = 8 * 1/2 = 4

New Rank (F) = PR(F) *W_t(F) = 4 * 1/2 = 2

New Rank (A) = PR(A) *W_t(A) = 9 * 1/3 = 3

Thus, it is evaluated from the above results that the ranking of search results has been modified to a great extant and the more relevant Web pages can be presented according to the above implementation. It is also results in the reduction of time complexity.

VII PROS AND CONS OF PROPOSED WORK

The various advantages of proposed work are as follows:

1. It helps in improving the quality of the search results i.e. relevant pages get higher ranking as compared to irrelevant results.
2. It provides an excellent opportunity for gaining insight into how a search engine is used and what the users' interests are since query log form a complete record of what users searched for in a given time frame
3. It helps in reduction of time complexity i.e., the time users spend for seeking out the required information is reduced significantly.
4. The shortcomings of traditional ranking methods are removed as the proposed algorithm is dependent on user's query words.

Some of the limitations of proposed work are:

1. The proposed clustering algorithm is only an offline procedure; it should be incremental for dynamic queries so that query clustering can be performed online.
2. Space complexity has increased due to the generation of large number of candidate sequences.
3. The pattern generation process leads to overhead as there are multiple number of scans in sequence database mining.

VIII CONCLUSION

A new approach based on log analysis is proposed for implementing interactive Web search. The most important feature is that ranking method is based on user's feedback to determine the relevance between Web pages and users' query words. The sequential patters are extracted from the document clicks of similar queries stored in a Query Cluster database, rather than from the contents of the retrieved documents. Furthermore, the rank of each Web page is calculated using the product of Page rank algorithm and weighted level of generated patterns. The proposed approach is dependent on user's query words and hence, the requirement of a user can be easily matched.

Since the results are based on analysis of the query logs, it shows that the ranking method is able to provide users relevant Web pages and reduce users' time on finding the required information from the search results list. The results obtained so far demonstrate that the proposed approach is quire promising in respect to improving the effectiveness of interactive web search engines.

In summary, the availability of large numbers of user logs provides new possibilities for search engines. In particular, it allows trends in user searching behavior to be spotted, thus helping builders of search engines and editors responsible for content to improve their system. The present study is only a first step in this direction.

REFERENCES

[1] Uniform Resource Identifiers (URI): Generic Syntax. http://www.rfcditor.org /rfc/rfc2396.txt,1998.
[2] Web Characterization Terminology & Definitions Sheet. http://www.w3.org/1999/05/WCA-terms/.W3C Working Draft 24-May-1999
[3] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan," Web Usage Mining: Discovery and applications of usage patterns from Web data", ACM, SIGKDD, volume 1 issue 2 pp. 12-23, 2000.
[4] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, "Grouping Web page reference into transactions for mining World Wide Web browsing patterns",1997.
[5] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, "Data preparation for mining World Wide Web browsing patterns",1999.
[6] F. Massegila, P.Poncelet, M.Teisseire, "Using data mining techniques on Web access logs to dynamically improve Hypertext structure",1999.
[7] L.Catledge and J. Pitkow, "Characterizing browsing behaviors on the world wide web", Computer Networks and ISDN Systems, 27(6),1995.
[8] Alex G. Buchner and Maurice D. Mulvenna, "Discovering Internet marketing intelligence through online analytical Web Usage mining", ACM SIGMOD Record, 27(4):54-61, December 1998.
[9] Common log file format. Retrieved June 02,2003 from http:/www.w3.org/Daemon/User/Config/Logging.html
[10] Extended log file format. Retrieved June 03,2003 from http:/www.w3.org/TR/WD-logfile.html
[11] CGI environment variables Retrieved May 15, 2003 from http://hoohoo.ncsa.uiuc.edu.cgi/env.html
[12] Peter, Pirolli, James Pitkow, and Ramana Rao, "Silk from a sow's ear. Extracting usable structures from the web", In CH1-96, Vancouver, 1996.
[13] G. Salton and M.J. McGill, "Introduction to Modern Information retrieval", McGraw-Hill. New York. 1983.
[14] E.Morphy, "Amazon Pushes 'Personalized Store for Every Customer", Ecommerce Times. September 28, 2001, http:/www.ecommerce.com/perl/story/13821.htm
[15] Amazon.com,www.amazon.com
[16] Google Inc. http://www.google.com/
[17] T. Springer, "Google LaunchesNewsService", PC World, September 23, 2002, http://www.computerworld.com/developmenttopics/websitemgmt/story/0,10801,00.html.
[18] DoubleClick's DART Technology,http://www/doulbleclick.com/dartinfo/.
[19] Amercia Online,www.aol.com
[20] eBay Inc., www.ebay.com
[21] E.Colet, "Using Data Mining to Detect Fraud in Auctions", DSStar, 2002.
[22] Yahoo!, Inc.www.yahoo.com
[23] D. Gusfield, "Inexact matching, sequence alignment, and dynamic programming", In Algorithm on Strings, Trees, and Sequences Computer Science and Computational Biology, Cambridge University Press, 1997.
[24] Dubes, R.C. and Jain, "Algorithms for Clustering Data", Prentice-Hall, Englewood Cliffs, NJ, 1988.
[25] Kulyukin, V.A., Hammond, K.J. and Burke, R.D., "Answering questions for an organization online, "In Proceedings of AAAI 98.532-538, 1998.

