International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0063
ISSN (Online): 2279-0071

International Journal of Software and Web Sciences (IJSWS)
www.iasir.net

Search Engine Optimization using Sequential Pattern Mining and Page Ranking

[1]Kulvinder Singh, [2]Yogesh Kumar
[1]Asst. Prof. [2] M.Tech. Research Scholar, Computer Science and Engineering Department,
University Institute of Engineering and Technology, Kurukshetra University, INDIA.

Abstract: *The premise of this paper is to scroll the various search engine optimization techniques which are used in solving problems associated while finding relevant information related to specific topic across the Internet. While carrying this review work the authors concentrated on various published research papers and finally concluded that Query Clustering, Sequential Pattern Mining and Page Ranking techniques are best suited for solving Search problems.*

Keywords: *Search engine, Information, Web mining.*

I. Introduction

The web has become a huge information repository. Having the vast volume of information on the Web, the task of finding relevant information related to a specific topic is becoming difficult. Many advanced Web searching techniques have been developed to reduce this problem and are being used in the commercial *web search engines* such as Google and Yahoo. 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 are many challenges in ranking the document's relevance based on user query

A. Architecture of Search Engine

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. A resource can be anything that has identity [1]. The examples include an electronic document, an image, a service (e.g. "today's weather report for Los Angeles") as well as a collection of other resources

The most important concept about the web is of course the resource that a server makes available to clients spread everywhere on the Internet; without any resource, the whole system would not have any sense. Indeed it would not be useful to make available some resources if nobody could access them, as well as clients without servers would be nonsense. When a resource is accessed by a client at a specific time and space, then it is called as resource manifestation.

B. Web Structure Mining (WSM)

The structure of a typical Web graph consists of Web pages as nodes, hyperlinks used to connecting between two related pages. WSM can be regarded as the process of discovering structure information from the Web. [2]

C. Classification of WSM

Hyperlinks: It is a structural unit that connects a Web page to unlike location, either within the same Web page or to a unlike Web page. A hyperlink that joins the different part of the same page is called an Intra-Document Hyperlink and a hyperlink that connects two different pages is called as Inter-Document Hyperlink. It has been a important body of work on hyperlink analysis, of which provides an up-to-date survey.

Web usage mining is achieved first by reporting visitors traffic information based on web server log files and other source of traffic data (as discussed above), Web server log files were used initially by the webmasters and system administrators for the purpose of "how much traffic they are getting, how many requests fail, and what kin of errors are being uncovered". Alternatively, Web server log files can be record and trace the visitors' on line behaviors. For example, after a few basic traffic studies, the log files be able to help us answer questions such as "from what on search engine visitors are coming? What is most and least popular on pages? Which browsers and operating systems are most commonly used by visitors?"

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 taken, it may be joint with other relational databases, on which the data mining techniques are implemented. Through some data mining technique for example association rules, path analysis, clustering and categorization, sequential analysis, visitors' behavior patterns are found and interpreted.

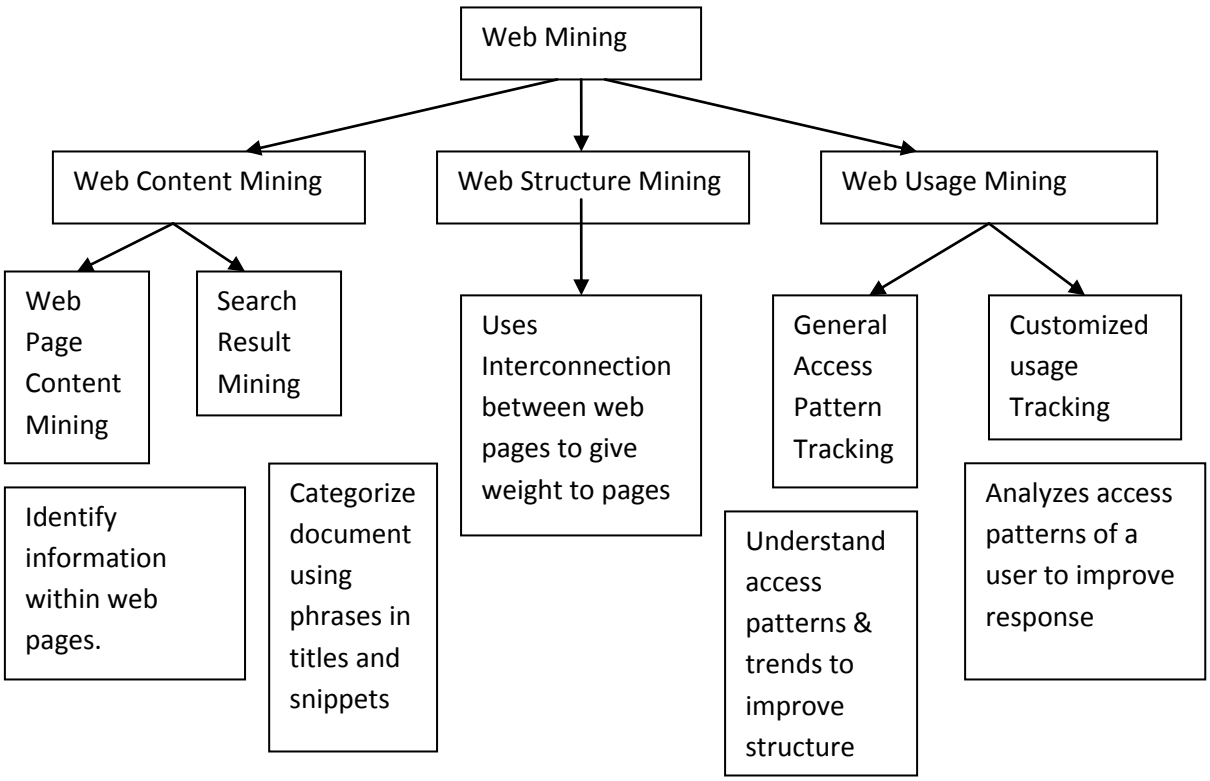

Figure 1.1: Taxonomy of Web Mining [3]

The above is the concise explanation of how Web usage is complete. Most complicated systems and techniques for discovery and analysis of patterns can be placed into three main categories: Preprocessing, Pattern Analysis Tools and Pattern Discovery Tools, as shown in Figure 1.2. These categories are explained below in detail.

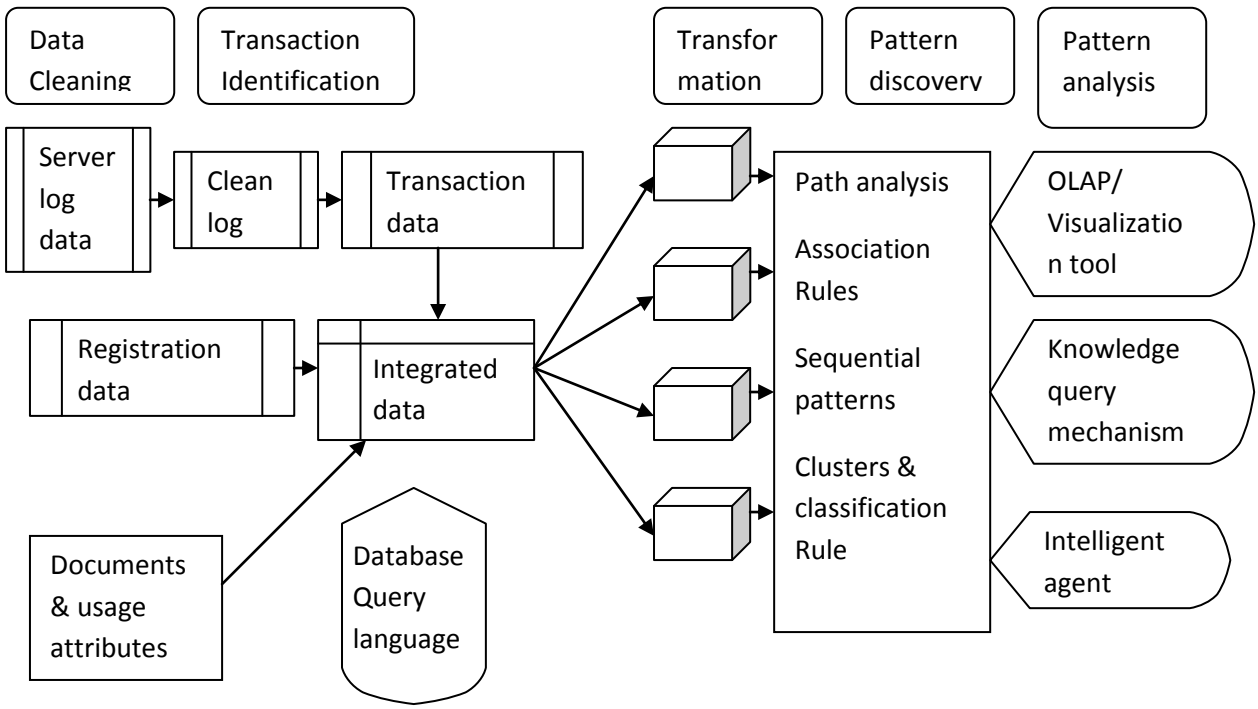

Figure 1.2: General Architecture for Web Usage Mining [4]

II. Previous Researches regarding Search Engine

Search engines are useful for finding information on the World Wide Web (WWW), such as Google, Yahoo and AltaVista. These general-purpose search engines are subject to low accuracy and low reporting. Manually-generated directories such as Yahoo! provide high-quality references, but cannot keep up with the Web's explosive growth. Although crawler-based search engines, like AltaVista, cover a larger fraction of the web,

their automatic indexing mechanisms often cause search results to be imprecise. It is thus difficult for a single search engine to offer both high reporting and high precision. This problem is exacerbated by the growth in Web size and by the increasing number of naive users of the Web who typically issues short (often, single word) queries to search engines. Topic-specific search engines often return higher-quality references than broad, general-purpose search engines for several reasons. First, specialized engines are often a front-end to a database of authoritative information that search engine spiders, which index the Web's HTML pages, cannot access. Second, specialized search engines often reflect the efforts of organizations, communities, or individual fanatics that are committed to providing and updating high quality information. Third, because of their narrow focus and smaller size, word-sense ambiguities and other linguistic barrier to high-precision search are ameliorated. The main stumbling block for a user who wants to utilize topic-specific search engines is: how do I find the appropriate specialized engine for any given query? Search.com offers a directory of specialized search engines, but it is up to the user to navigate the directory and choose the appropriate engine. A search engine of search engines is required. To build such an engine two questions have to be addressed: How can we build an index of high-quality, specialized search engines? And, given a query and a set of engines, how do we find the best engine for that query? In this paper, we focus on the latter problem, which is often referred to as the query routing problem.

Jing and Baluja [5] paper defined the Page Rank computation, a numerical weight was assigned to each image; this measures its relative importance to the other images was considered. The integration of visual signals in this process differs from the majority of large scale commercial search engines. After studying this paper image search has become a popular feature in many search engines, together with yahoo, Google, MSN etc., the majority of image searches use small, if any, image in formation to rank the images .Global features like color histograms and shape analysis, when used alone, are often too restrictive for the breadth of image types that need to be handled. A reliable measure of image similarity is crucial to good performance since this determines the underling graph structure. Worldwide features like color histograms and shape analysis, when applied alone, is also too restrictive for the breadth of image types that need to be handled. One method to reduce the computational cost is to pre cluster web images based using metadata such as anchor text, text, similarity or connectivity of the web pages.

Sharma and Dixit [6] described the World Wide Web is a global, large repository of text documents, pictures, multimedia and a lot of other items of information, referred to as information resources. To download a document, the crawler takes its seed URL and depending upon the host protocol and downloads the web document from web server. Keep the local collection fresh freshness of a collection can vary depending on the strategy used. Revisit frequency for a page based on its estimated change frequency. After studying this paper incremental crawler is used to In order to refresh its collection, a traditional crawler periodically replaces the old documents with the newly downloaded documents. It also replaces less important pages by new and more important pages. When the information contained in a document changes very frequently, the crawler downloads the document as often as possible and updates it into its database so that fresh information could be maintained for the potential users. It is the unique identifier for each document is called doc id.

Akansha and Krishna. [7] Discussed the web crawler was a module of a search engine that fetches data from a range of servers. Web crawlers were a necessary component to search engines running a web crawler is a challenging task. It was a time consuming process to collect data from a variety of sources around the world. A crawling module which fetches pages from web server is known as web crawler. A crawler can either be centrally managed or totally distributed is called parallel crawler. After studying this paper parallel crawlers provided the good result. It utilizes the memory of the machines and there is no disk access. MERCATOR is a scalable and extensible crawler, rolled into the ALTAVISTA search engine. Web server (WS) is the term web server means a computer program that is responsible for accepting HTTP requests from client's user agents such as web browsers and serves those HTTP responses along with optional data contents, which usually are web pages such as HTML documents and linked objects.

Raghavan and Paepcke [8] explored the idea of constructing and maintaining a large shared repository of web pages. It identified the functional modules and focused on the storage manager module, and traditional techniques for storage and indexing which meet the requirements of a web repository. The Google search engine computes the Page Rank of every web page by repeatedly observing the web's link structure. The warehouse receives web pages by a crawler, which is responsible for automatically finding new or customized pages on the web. At that time the database offers applications on access interface (API) so that they may efficiently access large numbers of up-to-date web pages. After studying this paper, streams are used in the repository needs to provide access to individual stored web pages, to large collections of pages, for indexing or data mining. The repository needs to handle a high rate of modifications. Repository strategy is used to avoid excessive conflicts between the update process and the applications accessing pages.

Heydon and Najork [9] described that Mercator was a scalable web crawler .It was extensible web crawler written entirely in Java. Scalable web crawlers were important components of many web services. It fetches tens

of millions of web documents. It provides the customizability. It's a mechanism for limiting pages which are crawled. Mercator supported for extensibility and customizability. In this paper it is found that it uses to rewind input stream to avoid the reading of a document over the network multiple times. It caches the document locally using an abstraction. In this small document was written in memory and larger documents were temporally written to a backing file. Mercator is an extensible web crawler. It has extended new functionality and fetching documents according to different new protocols. It was also used in the random walks to gather a sample of web pages. The sampled pages were used to measure the quality of search engine. It is easy in Mercator to configure the crawler for varying memory foot Prints.

Advantages of Search Engine Optimization

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.

III. Challenges in Search Optimization

1. Search engine give too many Web pages in output.
2. Users have to spend much time on finding their desired information from the long search result list.
3. The traditional ranking method is based on content-oriented and link-oriented approaches which give each Web page a score for ranking.
4. The ranking score is calculated by some sophisticated approaches and is independent of users' query words.
5. The relation between Web pages and the requirement of a user could not be completely matched. The most relevant Web pages to users' query words will not be shown at the top of the search results list.
6. The top ranked results for frequently occurring queries may not contain documents relevant to the users' search intent.
7. Fresh and relevant pages may not get high ranks for an underspecified query due to their freshness and to the large number of pages that go with the query, in spite of the fact that a large number of users have searched for parts of their content recently.

IV. Conclusion

This paper concludes that 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 significance between Web pages and users' query words. The sequential patterns are extracted from the document clicks of similar queries stored in a Query Cluster database, rather than from the contents of the retrieved documents. In addition, the rank of each Web page is calculated using the product of Page rank algorithm and weighted level of generated patterns.

This will show that the ranking method is able to provide users related Web pages and reduce users' time on finding the required information from the search results list

In summary, the availability of large numbers of user logs provides new possibilities for search engines. It allows react user searching behavior to be marked, thus helping builders of search engines and editors responsible for content to improve their system.

V. 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", 2000.

[4] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, "Grouping Web page reference into transactions for mining World Wide Web browsing patterns",1997.

[5] Yushi Jing, Shumeet Baluja "Page rank for image search" WWW 2008 / Referred Track: Rich Media Beijing, China. April 21-25, 2008.

[6] A.K. Sharma, Ashutosh Dixit " Self adjusting refresh time based architecture for incremental Web crawler" IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.12, December 2008.

[7] Akansha Singh, Krishna Kant Singh " Faster and efficient Web crawling with parallel migrating Web crawler" IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 11, May 2010.

[8] Jun Hirai Sri ram Raghavan, Hector Garcia-Molina Andreas Paepcke "Web Base: A repository of web pages" System Integration Technology Center, Toshiba Corp., 3-22 Katamachi, Fuchu, Tokyo 183-8512, Japan-1999

[9] Allan Heydo, Marc Najork "Mercator: A Scalable, Extensible Web Crawler" (Scalable, Extensible Web Crawler" World Wide Web Compaq systems Research C enter 130 Lytton Ave. Palo Alto, CA 94301(1999)