AN ANALYTICAL STUDY OF DATA MINING IN SEARCH ENGINE OPTIMIZATION TECHNIQUES

ASIT KUMAR MOHAPATRA

Research Scholar, Dept. of Computer Application,

Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal-Indore Road,
Madhya Pradesh, India,

Dr. Jitendra Seethlani

Research Guide, Dept. of Computer Application,

Sri Satya Sai University of Technology & Medical Sciences, Seh.ore, Bhopal Indore Road,
Madhya Pradesh, India.

ABSTRACT

Data mining is the application of sophisticated analysis to large amounts of data in order to discover new knowledge in the form of patterns, trends, and associations. For finding suitable information Search engines play an important role in retrieval of information on the web as they had become the entry point for accessing the web. Search engines analyse various aspects of the web page like its content and other attributes and display them accordingly .So it is important for the webmasters to develop and create those webpage's that fulfill the requirements of search engine but there is no such standard for ranking the webpage as it vary from one search engine to another. So webmasters uses different search engine optimization techniques like Keyword selection, Directory submission, Social bookmarking, Target market, Content, Keyword density in contents etc. to promote their webpage's. This paper focuses in analysing different search engine optimization techniques and finding those techniques that makes maximum impact in the ranking of the web page for this purpose researcher had used k-means cluster analysis for clustering various SEOT.

Keywords: Search engine optimization, Data mining, K-means cluster analysis, weka

INTRODUCTION

The growth of the Internet, its usage and dependency, leads towards various challenges. The Internet has opened up vast possibilities by opening the doors to Data control and access. It allows users to share their information and data through social networking site or simply by creating some Web Pages using different languages and technology. Search engines have a unique policy for indexing information in an efficient manner, and it is essential to optimize web-pages in a specific way to enhance their search ranking. Search engine optimization is about modifying the webpage accordingly to different parts of the website. These small modifications when viewed individually might make exponential improvements. Search engine are used to fetch the information through the World Wide Web and thus many challenges arises in the working of search engine. Search engine are used to fetch the list of websites with respect to the keywords query inputted by the user .Webmasters Optimizes their websites so as to get best position in the SERP. They use various optimization techniques to improve their rank in Search engine. Thus a need was found to analyse various search engine optimization techniques. Various researchers has worked on the subject and founded various Search engine optimization techniques to improve the rank of a website in search engine result page .In the past decade SEOT has drawn attention of webmasters since the use of search engine is preferred by the users.

So there is a need of proper classification of Search Engine Optimization Techniques (SEOT) that will help to improve the knowledge base and further leads to increase the efficiency of search engine. The purpose of study is to give guidelines and a better understanding to the persons engaged in the field of Search engine optimization and web developers that how to implement search engine optimization practices so as to improve their website ranking. Search engine optimization practices have grown increasingly important and this study suggest some techniques that have to be focused while promoting a website or a webpage. Data mining - also known as knowledge- discovery in databases (KDD) is process of extracting potentially useful information from raw data. A software engine can scan large amounts of data and automatically report interesting patterns without requiring human intervention. Other knowledge discovery technologies are Statistical Analysis, OLAP, Data Visualization, and Ad hoc queries.

REVIEW OF LITERATURE

Fu-Ming Hung and Jenn-Hwa Yang et al [11], present an intelligent search engine with semantic technologies. This survey has combine description logic inference system and digital library ontology to complete intelligent search engine.

Inamdar and Shinde et al [12], discussed agent based intelligent search engine system for web mining. Patrick Lambrix and Nahid Shahmehri and Niclas Wahllof et al [13], presents a search engine is described as one that tackles the problem of enhancing the precision and recall for retrieval of documents. There have been tested the system on small-scale databases with promising results.

Satya Sai Prakash et al [14], present architecture and design specifications for new generation search engines highlighting the need for intelligence and give a knowledge framework to capture intuition.

Dan Meng, Xu Huang et al, discussed an interactive intelligent search engine model based on user information preference [15]. This model can be an effective and useful way to realize the individuation information search for different user information preference.

Xiajiong Shen Yan Xu Junyang Yu Ke Zhang et al, forward an intelligent search engine where Information Retrieval model is found on formal context of FCA (formal concept analysis) and incorporates with a browsing mechanism. FCA is a useful way of supporting the flexible management of documents according to conceptual relation [16].

As popularity of WWW increases incrementally, millions of people use various search engines to discover information for the various web servers. But majority of users are interested only in few top listed result pages. Here comes the role of Search Engine optimization and hence promoting a website in search engine result page is a major task in website development and maintenance. Website ranking in search result strongly depends on how Search engine optimization (SEO) is implemented (Patil Swati , Pawar B.V., Patil Ajay S 2013).

Search engine becomes an integral part of everyone’s life to search information. The users rely on search engines to provide us right information at right time. To satisfy users need search engine must find and filter most relevant information matching a user query and display that information to the user (Cho J. And Roy S., 2004).

Search Engine Optimization (SEO) is a process/activity that relates with of optimizing websites/web-pages to achieve higher raking in the SERP. Major search engines rank individual web- pages or websites based on certain factors. The focused methodology used in Search engine optimization is to update both content and associated coding of the website/webpage to improve its visibility and ranking in organic searches made by the search engines (Rehman Khalil ur and Khan Ahmed Naeem Muhammad, 2013).

Search engine optimization increases the ranking of search results in the Internet marketing and they elaborated in their research that the rank of the motel sites and its bandwidth increased for Internet marketing after the implementation SEO techniques. The authors used several techniques of SEO to increase the bandwidth and ranking of search results including text title, label text, picture note, and HTML modification (H. L. Hsing, H. J. Chen, L. Me. Huang and H. H. Yi,2010).

SEARCH ENGINE ARCHITECTURE

Creating a search engine which scales even to today’s web presents many challenges. Fast crawling technology is needed to gather the web documents and keep them up to date. Storage space must be used efficiently to store indices and, optionally, the documents themselves. The indexing system must process hundreds of gigabytes of data efficiently. Queries must be handled quickly, at a rate of hundreds to thousands per second.

These tasks are becoming increasingly difficult as the Web grows. However, hardware performance and cost have improved dramatically to partially offset the difficulty. There are, however, several notable exceptions to this progress such as disk seek time and operating system robustness [2]. In designing Google, needs to consider both the rate of growth of the Web and technological changes. Google is designed to scale well to extremely large data sets. It makes efficient use of storage space to store the index. Its data structures are optimized for fast and efficient access. Further, expect that the cost to index and store text or HTML will eventually decline relative to the amount that will be available. This will result in favourable scaling properties for centralized systems like Google.

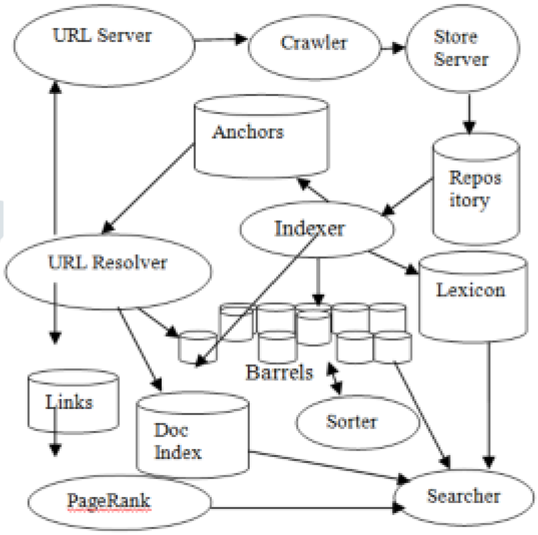

Fig. 1 Search Engine Architecture

The web crawling is done by several distributed crawlers. There is a URL server that sends lists of URLs to be fetched to the crawlers. The web pages that are fetched are then sent to the store server. The store server then compresses and stores the web pages into a repository. Every web page has an associated ID number called a doc ID which is assigned whenever a new URL is parsed out of a web page. The indexing function is performed by the indexer and the sorter. The indexer performs a number of functions.

It reads the repository; uncompressed the documents, and parses them. Each document is converted into a set of word occurrences called hits. The hits record the word, position in document, an approximation of font size, and capitalization. The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index. The indexer performs another important function. It parses out all the links in every web page and stores important information about them in an anchors file. This file contains enough information to determine where each link points from and to, and the text of the link. The URL resolver reads the anchors file and converts relative URLs into absolute URLs and in turn into doc IDs. It puts the anchor text into the forward index, associated with the doc ID that the anchor points to. It also generates a database of links which are pairs of doc IDs. The links database is used to compute Page Rank‟s for all the documents.

The sorter takes the barrels, which are sorted by doc ID and resorts them by word ID to generate the inverted index. This is done in place so that little temporary space is needed for this operation. The sorter also produces a list of word IDs and offsets into the inverted index. A program called Dump Lexicon takes this list together with the lexicon produced by the indexer and generates a new lexicon to be used by the searcher. The searcher is run by a web server and uses the lexicon built by Dump Lexicon together with the inverted index and the Page Ranks to answer queries.

A. Types of Search Engines

The entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes. Recommended font sizes are shown in Table 1.

B. Improvements of Search Engine

There are different ways to improve the performance of web search engines. Generally speaking, there are three main directions:

• Improving user interface on query input

• Using Filtering towards the query results

• Solving algorithms in web page spying and collecting, indexing, and output.

C. Basic Types of Search Tools

1) Crawler Based Search Engines

Crawler based search engines create their listings automatically. Computer programs „spiders‟ build them not by human selection. They are not organized by subject categories; a computer algorithm ranks all pages. Such kinds of search engines are huge and often retrieve a lot of information -- for complex searches it allows to search within the results of a previous search and enables you to refine search results. These types of search engines contain full text of the web pages the link to. So one can find pages by matching words in the pages one wants.

2) Human Powered Directories

These are built by human selection i.e. they depend on humans to create listings. They are organized into subject categories and subjects do classification of pages. Human powered directories never contain full text of the web page they link to. They are smaller than most search engines.

GENERAL SEARCH ENGINES

It includes the search engines Google yahoo etc. which provides a number of links when search the user for a query. It became a vast collection of information for these arches. It may not be containing the exact fact but it searches the query related all items which syntactically matches for the searched query. As far as users are concerned they need relevant and precise results.

A. Conventional Searching

Conventional searching helps the user to have the links of the searched query. It gives all the possible urls. In conventional searching it is not considering about the different meanings the words can have infarct it will show all the matches possible. By clicking or going through the links only have the clear picture about the query that what searching for. But it is not the case of the semantic search engines [3]. It is time consuming process if go through each and every links one by one. That may also happen in this type of search engines. While comparing this with the semantic search it is giving a difficult way to the user to get the result of the specified query.

The conventional search engines always provide the links for the user to go through to reach the results. It will also have the shot keys to search in the web, pictures, videos news, shopping etc… but the user will not get the answer for the query. In all these searches the search engine will provide the list of links by which the user can reach the destination. In conventional search engines it is not sure that the search thing is the same what is getting or the other possibilities of the same query.

SEARCH ENGINE OPTIMIZATION

Search engine optimization (SEO) refers to techniques that help your website rank higher in organic [5].

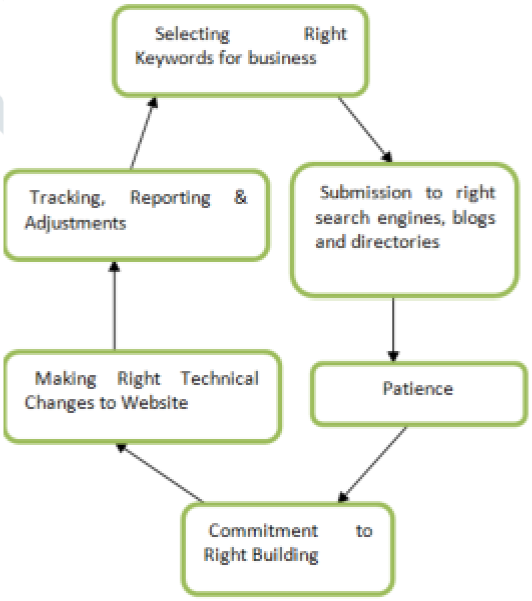
Selecting Right Keywords for business
Submission to right search engines, blogs and directories
Patience
Commitment to Right Building
Making Right Technical Changes to Website
Tracking, Reporting & Adjustments

Fig.2 Effective Search Engine Optimization

A. Search Engine Optimization Techniques

1) Directory Submission

Directory submission is one of the important techniques in SEO to create incoming links to a website through related page and category. A website is created and need to be rank to get good business results. Manually submission to directories is the best approach to rank your website. Internet directory is the platform on World Wide Web for information and links of many websites. Many directories are providing free service to website in directory [6]. To submit website in directories can produce web traffic for your website. This assist you to promote your business needs. The directory submission is used as SEO technique to promote your business.

2) Keyword Generation

Any search engine optimization method used keywords generation process. The keywords are necessary and most important part of SEO. The keywords are must be related to business [7]. Because related keywords boost website in short span of time. There are many online tools available to generate keywords relevant your needs like: Word tracker, Yahoo keyword selector tool, Google Ad words keyword tool and Thesaurus etc. By using these tools just put one word related your website like gamming. But only keywords are not providing assurance to popularity of website.

3) Link Exchanges

The link exchange is the method in SEO to place link on other websites and other websites place links on your websites means vice versa [8]. There are many types of link exchanges are used like: illustrate interest directly on web pages and other is that send email or discussion forums to show interest for link exchanges.

CONCLUSION

Data mining is implemented for finding some useful facts and patterns from different data sources. Use of data mining technique can help to understand and analyse data and information in proper way so that is will be helpful in different sectors. Search engine optimization is also a sector where there is a requirement of identifying some selective search engine optimization techniques from various SEOT so as to achieve a better rank in search engine result page. This study uses k-means clustering technique for clustering. The final results demonstrate that the proposed approach revealed those techniques that can help webmasters to achieve better rank in search engine result page. The purpose of study is to give guidelines and a better understanding to the persons engaged in the field of Search engine optimization and web developers that how to implement search engine optimization practices so as to improve their website ranking.

REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hyper textual Web search engine. In Proceedings of the Seventh WWW Conference, Brisbane, Australia, 1998.

[2] Grossan, B. "Search Engines: What they are, how they work, and practical suggestions for getting the most out of them," February1997.

[3] Koyoro Shadeo, Trends in web Based Search Engine „Journal of emerging trends in computing and information Sciences‟ Vol 3, No-6, June 2012, ISSN – 2079-8407.

[4] PSSE: Architecture for a Personalized Semantic Search Engine A. M. Riad, Hamdy K. Elminir, Mohamed Abu ElSoud, Sahar. F. Sabbeh. doi: 10.4156/aiss.vol2.issue1.9

[5] Bo Xing, Zhangxi Lin. The Impact of Search Engine Optimization on Online Advertising Market: The Eight International Conferences on Electronic Commerce (ICEC‟2006), pp. 519-529, ACM Electronic Commerce, 2006.

[6] Muhammad Akram, Search Engine Optimization Techniques Practiced in Organizations, "A Study of Four Organization", Journal of Computing, Vol-2, Issue-6, June-2010, ISSN- 2151-9617

[7] Dr. S. Sarvankumar, A New methodology for search engine optimization without getting sandboxed „International journal of Advanced research in computer and communication Engineering Vol 1, issues, Sept 2012, PP- 472-475.

[8] Mike Barus. "Link Exchange and One Way Links Using Web Directories," February 2009.

[9] „A web crawler design for data mining‟ Mike Thelwall Journal of information Science, 27 (5) 2001 PP. 321

[10] C. W. Cleverdon. The Cranfield tests on index language devices. In Aslib Proceedings, volume 19, pages 173-192, 1967. (Reprinted in Readings in Information Retrieval, K. Spärck- Jones and P. Willett, editors, Morgan Kaufmann, 1997).

[11] Fu-ing Huang et al. "Intelligent Search Engine with Semantic Technologies"

[12] S. A. Inamdar1 and G. N. Shinde "An Agent Based Intelligent Search Engine System for Web mining" Research, Reflections and Innovations in Integrating ICT in education 2008.

[13] Patrick Lambrix et al, "Dwebic: An Intelligent Search Engine based on Default Description Logics"-1997.

[14] K. Satya Sai Prakash and S. V. Raghavan “Intelligent Search Engine: Simulation to Implementation”, In the proceedings of 6th International conference on Information Integration and Web-based Applications and Services (iiWAS2004), pp. 203-212, September 27 - 29, 2004, Jakarta, Indonesia, ISBN 3-85403-183-01.

[15] Dan Meng, Xu Huang “An Interactive Intelligent Search Engine Model Research Based on User Information Preference”, 9th International Conference on Computer Science and Informatics, 2006 Proceedings, ISBN 978-90-78677-01-7.