Analytical Study of Information Retrieval techniques and Modified Model of Search Engine

Ms. Leena More[1] (Deshmukh)

1. Research Scholar, JJTU, Rajasthan and AP, JSPM's JIMS, Tathawade, Pune
(linadeshmukh@gmail.com)

ABSTRACT:

The concept of Information Retrieval is very vast and too many models of search engines are available in the market. In this research various information retrieval techniques used in search engine were studies and modified model of search engine were developed. In web mining most of the web search engines retrieve the documents or information first without knowing the meaning of the keyword and then ask for the relevant meaning of the keyword entered by the users. That means without understanding the exact meaning of keyword if it has synonymy or polysemy meanings. Search engine retrieve the documents as per its perception and then ask for did you mean? Due to that it takes more time to retrieve the relevant or quality documents. As in existing search engine there are too many problems by using previous techniques; there is need to develop, modify or combine more than existing algorithm. The proposed model; preference mining is based on existing model with few modifications. In this model first user's preference will be taken into account and then information gets retrieved. By using this model knowledge seekers will get relevant documents as per their interest within short period of time.

Keywords: synonymy, polysemy, Web crawlers, indexing, ontology

INTRODUCTION:

Web search is very different from normal information retrieval search of a printed document because of some factors like Bulk, Diversity, Growth, Dynamic, Demanding users, Duplication, Hyperlinks, Index Pages and Queries etc. Search engine do not search the web, they only search their databases.

Designing and building a good search engine is challenging task because of the scalability and performance. Search engines are huge databases of web pages as well as software packages for indexing and retrieving the pages that enable users to find information of interest to them.

The search engine databases of web pages are built and updated automatically by Web crawlers. Nobody is searching the entire Web. Instead one is only searching the database that has been compiled by the search engine. Huge database is searched using some kind of index and update their databases by using Web crawlers to find pages that have changed.

Web search engines either build directories like Yahoo! Or build full text indexes like Google to allow searches. There are also some meta-search engines that don't build and maintain their own databases but instead search the databases of other search engines.

Stemming algorithms can be quite complex and generally deal with prefixes as well as postfixes and must decide which affix is applied first. These algorithms are not perfect since they are based on heuristics.

Normally all search engines retrieve the documents or information first without knowing the meaning of the keyword and then ask for the relevant meaning of the keyword entered by the users. That means without understanding the exact meaning of keyword if it has synonymy or polysemy meanings. Search engine retrieve the documents as per its perception and then ask for did you mean? Due to that it takes more time to retrieve the relevant or quality documents.

The proposed model; **preference mining** is based on existing model with few modifications. In this model first user's preference will be taken into account and then information gets retrieved. By using this model knowledge seekers will get relevant documents as per their interest within short period of time.

LITERATURE REVIEW:

Devi et al. (2014), The PageRank and HITS algorithm give importance to links rather than the content of the pages. Both algorithms for ranking of web pages against the various parameters such as methodology, input parameters, relevancy of results and importance of the outcome, it is concluded that these techniques have limitations particularly in terms of time response, accuracy of results, importance of the outcome and relevancy of results.

Sharma D. K. and Sharma A. K. (2010), In this paper existing page ranking algorithm techniques have limitations particularly in terms of time response, accuracy of results, importance of the results and relevancy of results. An efficient web page ranking algorithm should meet out these challenges efficiently with compatibility with global standards of web technology.

Brin et al. (1998), Graph based algorithm based on link structure of web pages. Consider the back links in the rank calculations. Rank is calculated on the basis of the importance of pages. Results are computed at the indexing time not at the query time.

Kleinberg (1998), Rank is calculated by computing hub and authorities score of the pages in order of their relevance. Returned pages have high relevancy and importance with less efficiency and problem of topic drift.

Kim et al. (2002), This algorithm probabilistically estimates that clear semantics and the identified authoritative documents correspond better to human intuition. Well defined semantics with clear interpretation. Efficiently provide answer to quantitative bibliometric questions. Priori should be decided on the number of factors to model. Trades computational expense for the risk of getting stuck in local maxima.

Xing et al. (2004), Based on the calculation of the weight of the page with the consideration of the outgoing links, incoming links and title tag of the page at the time of searching. It gives higher accuracy in terms of ranking because it uses the content of the pages. It is based only on the popularity of the web page.

BaezaYates et al. (2004), This algorithm ranks the page by providing different weights based on three attributes i.e. relative position in page, tag where link is contained & length of anchor text. It has less efficiency with reference to precision of the search engine. Relative position was not so effective, indicating that the logical position not always matches the physical position

Fujimura et al. (2005), Use of the adjacency matrix, constructed from agent to object link not by page to page link. Three vectors i.e. hub, authority and reputation are needed for score calculation of the blog. Useful for ranking of blog as well as web pages because input and output links are not considered in the algorithm. Specifically suited for blog ranking.

Bidoki et al. (2007), Based on reinforcement learning which consider the logarithmic distance between the pages. Algorithm consider real user by which pages can be found very quickly with high quality. A large calculation for distance vector is needed, if new page inserted between the two pages.

Jiang et al. (2008), Visitor time is used for ranking. Use of sequential clicking for sequence vector calculation with the uses of random surfing model. Useful when two pages have the same link structure but different contents. It does work efficiently when the server log is not present.

Jie et al.(2008), The algorithm is based on the analysis of tag heat on social annotation web. Ranking results are very exact and new information resources are indexed more effectively. Co-occurrence factor of tag is not considered which may influence the weight of the tag.

Lamberti et al. (2009), Ranking of web pages for semantic search engine. It uses the information extracted from the queries of the user and annotated resources. Effectively manage the search page. Ranking task is less complex. In this ranking algorithm every page is to be annotated with respect to some ontology, which is the very tough task.

Lee et al. (2009), Individual models are generated from training queries. A new query ranked according to the combined weighted score of these models. It gives the results for user's query as well as results for similar type of query. Limited numbers of characteristics are used to calculate the similarity.

Chakrabarti (2002) and Manning et al. (2008) provides detailed coverage of Web crawling, ranking techniques, and mining techniques related to information retrieval such as text classification and clustering.

Brin and Page (1998) describe the anatomy of the Google search engine, including the PageRank technique, while a hubs- and authorities based ranking technique called HITS is described by **Kleinbeg (1999)**. **Bharat and Henzinger (1998)** present a refinement of the HITS ranking technique. These techniques as well as other popularity based ranking techniques and techniques to avoid search engine spamming are described in detail in **Chakrabarti (2002). Chakrabarti et al. (1999)** addresses focused crawling of the Web to find pages related to a specific topic. He provides a survey of Web resource discovery.

PROBLEMS OF EXISTING SEARCH ENGINE:

By using this technique the major drawback is that it assigns a measure of popularity that does not take query keywords into account. Ie Page Rank's algorithm focuses on the importance of a page rather than on its relevancy given the user query. Page Rank is calculated independently of a user query so the result served on the first screen may not be the most relevant; it may be the one with highest Page Rank amongst the pages retrieved. Search results are based on the literal (keywords, tags, meta data) things but not on meaning.

Also many SEO (Search Engine Optimization) industries can improve/manipulate the Page Rank of pages on the web using different techniques such as adding more keywords through META tags, trying to influence links within their web pages etc.

New pages have less page rank and they take much time to be getting listed and gain high ranks. So Page Rank does not deal with new pages fairly since it makes high Page Rank pages even more popular by serving them at the top of the results. Thus the rich get richer and poor get poorer.

Normally all search engines retrieve the documents or information first without knowing the meaning of the keyword and then ask for the relevant meaning of the keyword entered by the users. That means without understanding the exact meaning of keyword if it has synonymy or polysemy meanings. Search engine retrieve the documents as per its perception and then ask for did you mean? Due to that it takes more time to retrieve the relevant or quality documents.

Stemming algorithms can be quite complex and generally deal with prefixes as well as postfixes and must decide which affix is applied first. These algorithms are not perfect since they are based on heuristics.

OBJECTIVES:

Review of previous literature shows that there are many problems involved when retrieving information as per users need.

To overcome limitations of previously used techniques I designed some objectives of my research which are as below:

1. To study the various techniques used for information retrieval and their limitations.
2. To identify the factors due to which information retrieval is not up to the mark in existing techniques.
3. Designing and development of preference mining model/algorithm

THE QUALITY OF SEARCH RESULTS:

The results form a search engine ideally should satisfy the some quality requirements.

Precision: Only relevant documents should be returned.

Recall: All the relevant documents should be returned.

Ranking: A ranking of the documents providing some indication of the relative ranking of the results should be returned.

First Screen: The first page of the results should include the most relevant results.

Speed: Results should be provided quickly since users have little patience.

WORKING OF EXISTING SEARCH ENGINE:

The search engine carries out a variety of tasks. These include:

Collecting information: It would normally collect web pages or information about them by web crawling.

Evaluating and categorizing information: when web pages are submitted to directory, it may necessary to evaluate and decide whether a submitted page should be selected. It may be also necessary to categorize information based on some ontology used by search engine.

Creating database and indexes: the information collected needs tobe stored either in a database or some kind of file system. Indexes must be created so that the information may be searched efficiently.

Computing ranks of the web documents: different methods are used to determine the rank of each page retrieved in response to user query. The information used may include frequency of keywords, value of in-links and out-links from the page and frequency of use of the page.

Checking queries and executing them: for spelling errors and whether words in the query are recognizable is checked.

Presenting results: it determine what results to present and how to display them.

Profiling the users: to improve search performance, it carry out user profiling that deals with the way users use search engines.

ARCHITECTURE OF SEARCH ENGINE:

No two search engines are exactly same in terms of size, indexing techniques, page ranking algorithms, or speed of search. It has been found that if the same query is posed to two search engines, the results are often different.

A typical search engine architecture is as in Figure 1 consists of many components including three major components.

The crawler and the indexer: It collects pages from the web, creates and maintains the index.

The user interface: It allows users to submit queries and enables result presentation.

The database and the query server: It stores the information about the web pages and processes the query and returns results.

All search engines essentially include a crawler, an indexer and a query server; although the algorithms used and its quality may vary significantly.

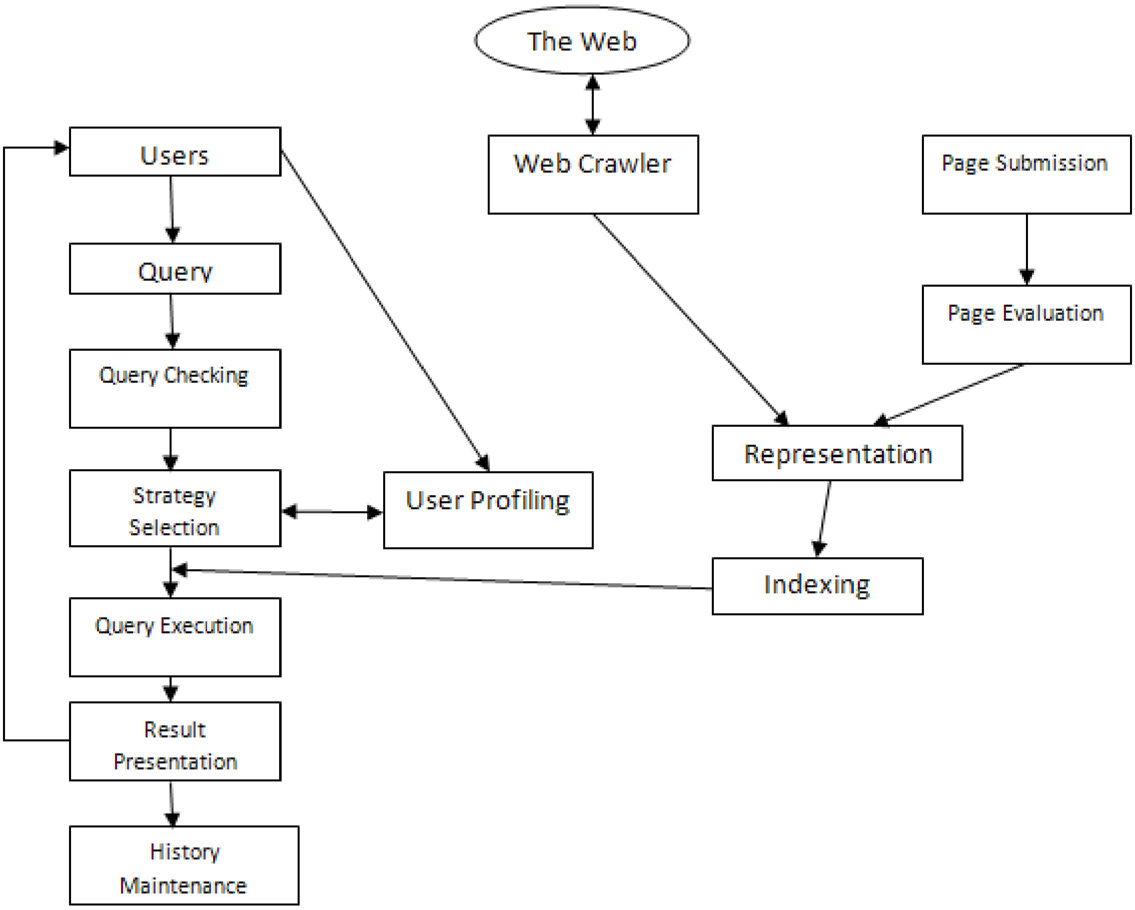
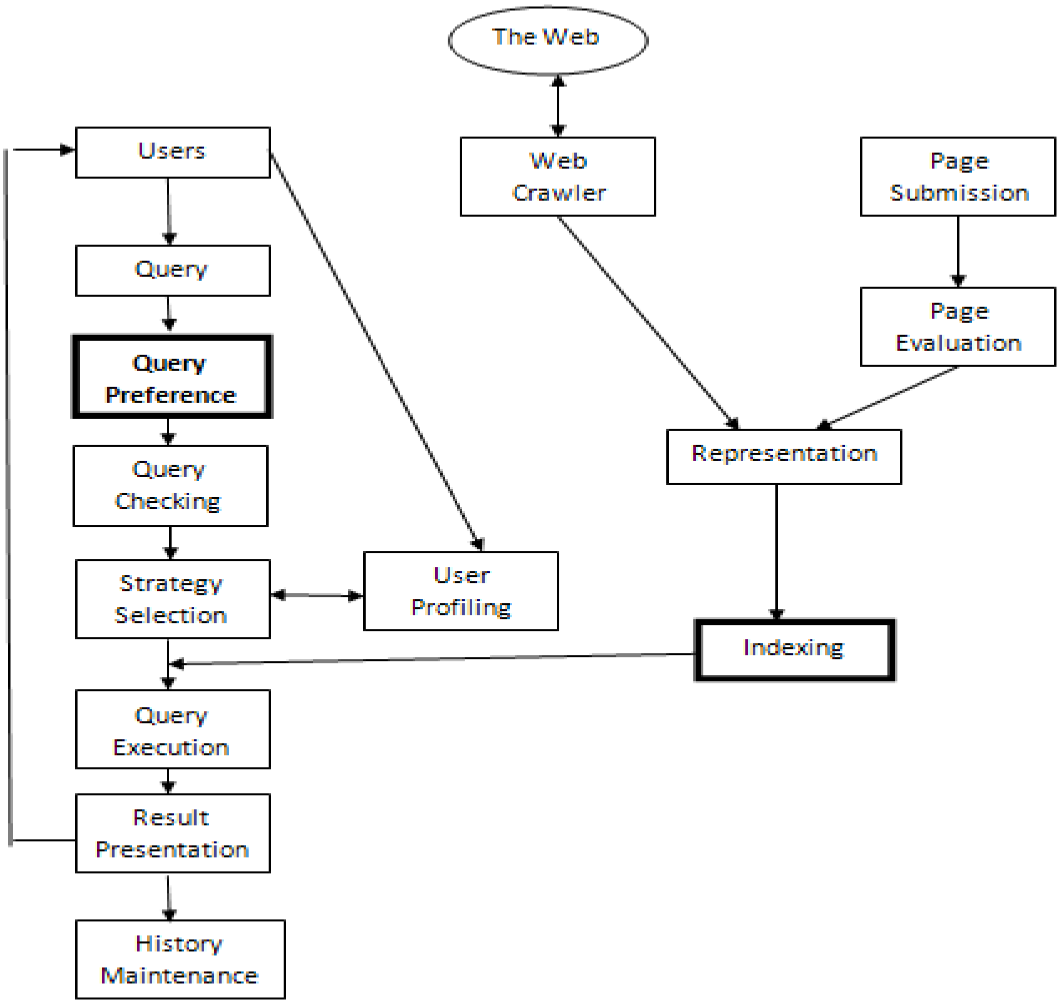


Figure 1 : Typical architecture of a search engine

PROPOSED SEARCH ENGINE:

Normally all search engines retrieve the documents or information first without knowing the meaning of the keyword and then ask for the relevant meaning of the keyword entered by the users. That means without understanding the exact meaning of keyword if it has synonymy or polysemy meanings. Search engine retrieve the documents as per its perception and then ask for did you mean? Due to that it takes more time to retrieve the relevant or quality documents. Also Stemming algorithms are not perfect since they are based on heuristics.

The proposed model; **preference mining** is based on existing model with few modifications. In this model first user's preference will be taken into account and then information gets retrieved. By using this model knowledge seekers will get relevant documents as per their interest within short period of time.



CONCLUSION:

All search engines retrieve the documents or information first without knowing the meaning of the keyword entered by the user and then asks for the relevant meaning of the keyword. That means without understanding the exact meaning of keyword if it has synonymy or polysemy meanings. Search engine retrieve the documents as per its perception and then ask for did you mean? Due to that it takes more time to retrieve the relevant or quality documents.

In proposed model first user's preference will be taken into account and then information gets retrieved. By using this model knowledge seekers will get relevant documents as per their interest within short period of time.

REFERENCES:

[1] Abraham Silberschatz, Henry Korth, S. Sudarshan (2011), Database System Concepts, McGraw Hill.
[2] Asmaa Benghabrit, Brahim Ouhbi, Hicham Behja, Bouchra Frikh (2013), Statistical and Semantic Feature Selection for Text Clustering, Journal of Intelligent Computing, volume-4, No.2, PP. 69-79
[3] Dilip Kumar Sharma, A. K. Sharma (2010), A Comparative Analysis of Web Page Ranking Algorithms, (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, PP 2670-2676
[4] G.K. Gupta (2013), Introduction to Data Mining with case studies, PHI.
[5] Jaiwei Han, Micheline Kamber (2006-2009), Data Mining Concepts and Techniques, Elsevier.
[6] K. S. Kuppusamy and G Aghila (2011), Web Content Mining tools: A Comparative Study, Volume 4, No. 2, PP. 485-488
[7] Michael J.A. Berry, Gordon S. Linoff (2010), Data Mining Techniques, Wiley.
[8] Peddi Kishor and Yohan Kasarla (2013), Research issues in data stream Association Rule Mining, IJCSKE, vol. 7, No. 1, PP. 16-26
[9] Pooja Devi, Ashlesha Gupta, Ashutosh Dixit [2014], Comparative Study of HITS and PageRank Link based Ranking Algorithms, International Journal of advanced Research in Computer and Communication Engineering Vol. 3, PP 5749-5754, Issue 2
[10] Prasad J C and K S M Panicker (2013), String Searching Algorithm Implementation – Performance study with two cluster configuration, IJWCS, volume 3, No. 2, PP. 83-87
[11] R. A. Baeza-Yates and B.A. Ribeiro-Neto (1999), Modern Information Retrieval, ACM Press/Addison-Wesley.
[12] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg (1999), "Mining the Web's Link Structure", Computer, 32(8), PP.60–67.
[13] V. Bharanipriya, Kamakshi Prasad (2011), Web Content Mining tools: A Comparative Study, Volume 4, No. 1, PP. 211-215

BIOGRAPHICAL NOTES:

Prof. Leena More (Deshmukh), has obtained Master of Computer Applications Degree from Shivaji University, Kolhapur in 2002. She is a Research scholar of JJT University, Rajasthan and also working as a AP in Department of MCA at JSPM's JIMS. She has more than 10 years of industrial and teaching experience for PG course. She has guided more than 90 academic projects at PG level. She has attended several National and International seminars and conferences and presented research papers. Her research interest includes Web Mining and Web Search Engine.