Study of Information Retrieval techniques of Search Engine with Special reference to different Institutions in Pune Region

Ms. Leena More (Deshmukh)

Research Scholar, JJT University, Rajasthan and Asst. Prof., JSPM's JIMS, Tathawade

linadeshmukh@gmail.com

Dr. Manik Kadam

Research Guide,

Allana Institute of Management, Pune

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 studied and reviews from different institutions / organizations about information retrieval has been taken into consideration. By this study many organizations knowledge seekers agreed with two major problems of efficiency and effectiveness due to retrieval without understanding meaning and stemming problem respectively. In web mining most of the web search engines retrieve the documents or information first without knowing the exact meaning of the keyword for synonymy or polysemy and then ask for the relevant meaning of the keyword entered by the users. That is it retrieve the documents as per its perception and then ask for did you mean? Secondly at the time of stemming some irrelevant stem has been taken into account and irrelevant documents gets retrieved. Due to this efficiency and effectiveness of search engine gets degraded.

Keywords: *stemming, polysemy, synonymy, Indexing, Text Transformation and Text Acquisition,* crawler

I. 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.

II. 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.

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.

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.

III. Problem Statement

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.

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.

IV. 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 search engines limitations in different institutions point of view for information retrieval through Web Search Engine.
2. To identify the factors due to which information retrieval is not up to the mark in existing techniques of search engine.

V. Hypothesis

H1: All Institutions always use the "Did You Mean?" Link for searching.

H2: Search Engine does not face the problem of irrelevant documents due to wrong stemming

H3: Search Engine provides duplicated web pages

H4: Search Engine provides required documents on first page only

H5: Search Engine provides good quality results

Research Methodology:

Research Design: Descriptive Research

Data Source: Primary as well as Secondary

Research Instrument: Questionnaire, F to F

Sample Unit: Researchers, Faculties and Students from different Institutions and Organizations

Sampling Method: Random Sampling

Sample Size: 53

Analysis and Findings:

The data collected through questionnaire were coded and tabulated keeping in context the objective of the study. Institution type wise distribution of respondents is given below-

Institution type wise distribution

Institution Type	**Frequency**	**Percentage**
Management	34	64.2
Engineering	7	13.2
IT Industry	12	22.6
Total	**53**	**100**

FINDINGS:

Q.1) How many times you go through Suggested links at bottom for searching relevant documents?

Institution type wise Response

	Yes		No	
Institution Type	**Frequency**	**Percentage**	**Frequency**	**Percentage**
Management	7	20.6	27	79.4
Engineering	2	28.6	5	71.4
IT Industry	2	16.7	10	83.3
Total	**11**	20.8	**42**	79.2

H0: All Institutions always use the links "Did You Mean?" for searching.

Calculation of Chi-Square

Institution Type	**Frequency**	**Percentage**	**fe**	**(fo-fe)2**	**(fo-fe)2/fe**
Management	7	63.6	3.7	11.1	3.0
Engineering	2	18.2	3.7	2.8	0.8
IT Industry	2	18.2	3.7	2.8	0.8
Total	**11**	100.0	11.0	16.7	4.5

Result: χ^2 at d.f of 2 and level of significance of 5% is which is 5.991 which is more than calculated value 4.5, therefore null hypothesis is accepted. It is proved that all types of institutions always use the links "Did you mean?" for searching

Q.2) How many times you got irrelevant documents due to wrong stem (ex. searched for copper and got results on copper, cope and cop)?

Institution type wise Response

	Yes		No	
Institution Type	**Frequency**	**Percentage**	**Frequency**	**Percentage**
Management	4	11.8	30	88.2
Engineering	0	0.0	7	100.0
IT Industry	0	0.0	12	100.0
Total	**4**	7.5	**49**	92.5

H0: Search Engine does not face the problem of irrelevant documents due to wrong stemming

Calculation of Chi-Square

Institution Type	**Frequency**	**Percentage**	**fe**	**(fo-fe)2**	**(fo-fe)2/fe**
Management	4	100.0	1.3	7.1	5.3
Engineering	0	0.0	1.3	1.8	1.3
IT Industry	0	0.0	1.3	1.8	1.3
Total	**4**	100.0	4.0	10.7	8.0

Result: χ^2 at d.f of 2 and level of significance of 5% is which is 5.991 which is less than calculated value 8.0, therefore null hypothesis is rejected. It is proved that all types of institutions always face the problem of irrelevant documents due to wrong stemming

Q.3) How many times you get duplicated web pages?

Institution type wise Response

	Yes		No	
Institution Type	**Frequency**	**Percentage**	**Frequency**	**Percentage**
Management	5	14.7	29	85.3
Engineering	2	28.6	5	71.4
IT Industry	2	16.7	10	83.3
Total	**9**	17.0	**44**	83.0

H0: Search Engine provides duplicated web pages

Calculation of Chi-Square

Institution Type	Frequency	Percentage	fe	(fo-fe)2	(fo-fe)2/fe
Management	5	55.6	3.0	4.0	1.3
Engineering	2	22.2	3.0	1.0	0.3
IT Industry	2	22.2	3.0	1.0	0.3
Total	**9**	100.0	9.0	6.0	2.0

Result: χ^2 at d.f of 2 and level of significance of 5% is which is 5.991 which is more than calculated value 2.0, therefore null hypothesis is accepted. It is proved that Search Engine provides duplicated web pages

Q.4) How many times you get required documents on first page?

Institution type wise Response

	Yes		No	
Institution Type	**Frequency**	**Percentage**	**Frequency**	**Percentage**
Management	27	79.4	7	20.6
Engineering	4	57.1	3	42.9
IT Industry	8	66.7	4	33.3
Total	**39**	73.6	**14**	26.4

H0: Search Engine provides required documents on first page only

Calculation of Chi-Square

Institution Type	Frequency	Percentage	fe	(fo-fe)2	(fo-fe)2/fe
Management	27	69.2	13.0	196.0	15.1
Engineering	4	10.3	13.0	81.0	6.2
IT Industry	8	20.5	13.0	25.0	1.9
Total	**39**	100.0	39.0	302.0	23.2

Result: χ^2 at d.f of 2 and level of significance of 5% is which is 5.991 which is less than calculated value 23.2, therefore null hypothesis is rejected. It is proved that Search Engine never provides required documents on first page

Q.5) How many times search engine provide Good Quality knowledge/search?

Institution type wise Response

	Yes		No	
Institution Type	**Frequency**	**Percentage**	**Frequency**	**Percentage**
Management	34	100.0	0	0.0
Engineering	7	100.0	0	0.0
IT Industry	10	83.3	2	16.7
Total	**51**	96.2	**2**	3.8

H0: Search Engine provides good quality results

Calculation of Chi-Square

Institution Type	Frequency	Percentage	fe	(fo-fe)2	(fo-fe)2/fe
Management	34	66.7	17.0	289.0	17.0
Engineering	7	13.7	17.0	100.0	5.9
IT Industry	10	19.6	17.0	49.0	2.9
Total	**51**	100.0	51.0	438.0	25.8

Result: χ^2 at d.f of 2 and level of significance of 5% is which is 5.991 which is less than calculated value 25.8, therefore null hypothesis is rejected. It is proved that Search Engine never provides Good Quality results.

VI. 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. Also stemming problem faced by users will be overcome through it. By using this model knowledge seekers will get relevant documents as per their interest within short period of time.

VII. Suggestions and Recommendations

Modification of existing search engine is required to improve efficiency and effectiveness of Web search engine.

- In this model first user's preference has to be taken into account and then information gets retrieved.
- Secondly as there is a stemming problem categorization of terms with term meaning should be done.
- Measure of popularity does not take query keywords into account. So Algorithm must focus on the importance of its relevancy given the user query rather than on page.
- New pages should gain high ranks as per their relevancy and should take less time to be getting listed.
- It should overcome from the problem of Google Bombing (Page Rank Manipulation).

VIII. References

I. *Abraham Silberschatz, Henry Korth, S. Sudarshan (2011), Database System Concepts, McGraw Hill.*

II. *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*

III. *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*

IV. *G.K. Gupta (2013), Introduction to Data Mining with case studies, PHI.*

V. *Jaiwei Han, Micheline Kamber (2006-2009), Data Mining Concepts and Techniques, Elsevier.*

VI. *K. S. Kuppusamy and G Aghila (2011), Web Content Mining tools: A Comparative Study, Volume 4, No. 2, PP. 485-488*

VII. *Michael J.A. Berry, Gordon S. Linoff (2010), Data Mining Techniques, Wiley.*

VIII. *Peddi Kishor and Yohan Kasarla (2013), Research issues in data stream Association Rule Mining, IJCSKE, vol. 7, No. 1, PP. 16-26*

IX. *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*

X. *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*

XI. *R. A. Baeza-Yates and B.A. Ribeiro-Neto (1999), Modern Information Retrieval, ACM Press/Addison-Wesley.*

XII. *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.*

XIII. *V. Bharanipriya, Kamakshi Prasad (2011), Web Content Mining tools: A Comparative Study, Volume 4, No. 1, PP. 211-215*

IX. *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.