Original Article

Bulletin of Pure and Applied Sciences.
Vol. 37E (Math & Stat.), No.2, 2018.P.311-320
Print version ISSN 0970 6577
Online version ISSN 2320 3226
DOI: 10.5958/2320-3226.2018.00034.6

REMOVING NON-RELEVANT LINKS FROM TOP SEARCH RESULTS USING FEATURE SCORE COMPUTATION

Swati P. Patil[1,*], B.V. Pawar[2]

Author Affiliation:
[1]Department of Computer Science, SSVPS's Science College, Dhule, Maharashtra 424005, India
E-mail: swatippatil28@gmail.com
B.V. Pawar
[2]School of Computer Sciences, North Maharashtra University, Jalgaon, Maharashtra 425001, India
E-mail: bvpawar@hotmail.com

***Corresponding Author:**
Swati P. Patil, Department of Computer Science, SSVPS's Science College, Dhule, Maharashtra 424005, India

E-mail: swatippatil28@gmail.com

Received on 09.02.2018, Accepted on 19.07.2018

Abstract

To promote website in search engine rankings in order to get better visibility and more traffic, search engine optimizers have to follow legal ways but some search engine optimizers manipulate the web pages and boost the irrelevant pages in search results and this leads to practice the problem of web spam. These malicious pages reduce the performance of the search engine. We innovates a method to override the irrelevant search results. In this paper a new approach is designed and developed to move non-deserving pages downward from list of results returned by search engine. The first step was to determine the importance of different page features used for the ranking of a page in search results. Based on this information the devised system identified the features in a page and assigned some weight to each feature. The user query i.e. search keyword, precision and ranking factor are used to calculate precision score of a page. This score is used to predict ranking position for each page. A comparison was made between original ranking positions and predicted ranking position of irrelevant web pages. From analysis of results of Google, accuracy of search results improved from 88% to 99% for the corpus compiled now.

Keywords: Feature, Search Engine Optimization, Rank.

1. INTRODUCTION

The number of websites and internet users is increasing continuously. Search engines help the user to find relevant information effectively and efficiently. Now a day Google is the most popular search engine. Google is not only a valuable search tool but a necessary one for wide range of applications [1]. In this research work Google's search result are taken to find out deserving and non deserving sites related to a given query in top search results.

How a search engine decides which pages are best matches and what order results should be shown as it varies widely from one search engine to another [2]. The search engine algorithms are kept secret by search engines.

According to research to determine the most-relevant pages, a search engine selects a set of pages that contain some or all of the query terms and computes a page score for each page. Finally, a list of pages, sorted by their score, is returned to the user[3]. This score is calculated using the properties of the candidate pages. These properties are called features. The important page features are used to rank the search results.

The search results getting top ranking are considered more important. In the case of the majority of web searches, approximately 80% visit no more than top 10 to 20 result pages [4]. Statistical data collected in 2013 indicates that Page 1 results garnered 92 percent of all traffic from the average search [5]. More visitors indicate more business. To place website in top search results is one of the strongest contributions to commercial websites success [6]. As a result, it becomes ever more difficult for websites to keep position in top search results among all the other competing sites. One direct way to achieve better ranking is to improve the quality of web pages. But this approach requires more money, time and resources. Instead of that search engine optimizers find a short cut to achieve this goal. They make use of the search engine optimization process (SEO). SEO is the process which improves quality and volume of web pages via natural search results [7]. SEO are preferred rather than Internet advertisement because of its lower cost [8]. Some search engine optimizers misguide the search engine by designing the website using SEO techniques in unethical ways. They manipulate the search engine ranking so that irrelevant sites are placed in top search results. To raise ranking position web attackers manipulate the web page by using misguiding keywords, keyword stuffing in the text, link farm and create doorway pages [9].

If legal SEO techniques are used then it improves the ranking of deserving sites and is beneficial to the user. But unethical SEO techniques mislead search engine. It pushed undeserving sites on top list of search engine which leads to the problem of large web spamming. Web search engines continuously update their algorithms and invest a lot of money to find solution over it. But still they face the challenging problem in maintaining the quality of search engines [10].

In this work, we determined the importance of different page features is determined for the ranking of a relevant web page in search results. The devised system has identified such features present in a web page and assigned these features have been assigned a weight in such a way that the score of relevant pages improved and relevant web sites move upwards in search results.

The remainder of the paper is organized as follows. Section II covers the related work done. Section III shows the methodology used. Section IV discusses the results obtained using feature score computation algorithm. Finally section V ends with the conclusion.

2. RELATED WORK

In last few years research work of search engine optimization has spread widely. Patil S. P. et. al. have discussed different SEO techniques used by search engine optimizers. This paper emphasizes the white hat SEO and black hat SEO techniques with its merits and de-merits [11]. Cui M. and Hu S. introduces new website building concept for construction of search engine optimization. In this paper features of search engine are explained and proposed search engine optimization tools, strategies and methods are presented, and analyzed the new thought that e-commerce sites with the search engine do the effective website promotion[12]. Yunfeng M. analyzed the impact of receiving and recording of search engines and ranking rules to the get understanding of the features of search engine algorithms commonly used and proposes the optimization tactics for the development of a website [13]. Zhu C. and Wu G. build a system which automatically crawl all factors of 200 thousand web pages. They follow the reverse engineering approach to study and analyze the key factors which influence the ranking result. Based on top 20 results of Google search result pages made their content analysis and derived from them the top five factors for search engine optimization [14]. Wang F. et. al. analyzed the impact of SEO techniques on the effectiveness of SEO to figure out which technique strategy is most effective, and furthermore, test the possible influence of SEO techniques on Page Interest. This paper attempts to evaluate the techniques of SEO by means of the third-party measuring tool based on the data collected from 116 websites [15]. Fiefei X. and Guangnian Z. developed a system called SEOAdvisor which is auto analyzing and verifying search engine algorithm and which is to made using statistical principles and their comparisons are suggest results. The system can predict and verify ranking algorithm used in popular search engines via capturing and comparing web pages listed on the top of search engine results page (SERP) automatically. Using this system search engine optimizer gets high ranking via optimizing website [16] Shi J. et. al. studies the university journal website and finds out ways to promote these websites in search engine results. In this paper, the features of university journal websites are studied, defined SEO strategies from those aspects like directory structure, keyword strategy, URL pseudo-static, code optimization and inbound links. [17]. Su J. et. al. focused on the Google ranking algorithm and design, implement, and evaluate a ranking system to systematically validate assumptions about Google ranking algorithm. This paper demonstrates that linear learning models, coupled with a recursive partitioning ranking scheme, are capable of reverse engineering Google's ranking algorithm with high accuracy. The system correctly predicts 7 out of the top 10 pages for 78% of evaluated keywords and can correctly predict 9 or more pages out of the top 10 ones for 77% of search terms. This work provide guidelines for search engine optimizers and webmasters to optimize their web pages, validate or disapprove new ranking features, and evaluate search engine ranking results for possible ranking bias[18]. Somani A. and Suman U. characterized some commonly used black hat SEO techniques, and proposed a new way to counter those techniques using link based spam detection combined with the page rank algorithm. This technique helps us to discover target page and trace down the entire graph responsible for spreading spam [19]. Boris K. and Marijana V. suggest major factors for good ranking position in all major search engines. This paper discusses advantages and disadvantages of search engine optimization [7].

3. METHODOLOGY

In this section detailed techniques are introduced which are used to include important features. First it is discussed why top pages of different search engines are selected. Then, those features which are selected are described. Then it is discussed how these features are used to compute score of a page. Finally, the report is presented on the rankings that major search engines produced for these pages and then conclusions could be draw about evaluation of search results

A. Selection of pages

In the present study the first 10 random queries were chosen. These queries are from different areas. These queries are fired on major 6 search engines like Google, Yahoo, Bing, Ask, Gigablast and About. A sample corpus was built by downloading several documents from the web using predefined set of queries. All search result pages with their URL were stored for future processing. Spink A. and et. al. claim that results which are not among top 10 are nearly invisible to general user [20]. Even the tendency of the user is to look often only at results set that can be seen without scrolling [21]. With this concept in mind top 20 pages were selected returned by each query for each search engine. Total 1400 results were selected.

B. Feature Selection

Each page was manually observed with its source code and disclosed 52 different features used as shown in Table I. From these an optimal subset of features which are most important in ranking of pages was identified. A feature is a property of a web page such as number of links pointing to other pages, frequency or location of keywords or presence of keywords in title tag, meta description tag H1 tag, anchor text etc.[3].

C. Extraction of Important Features

To find out most important features, at first html source code were examined and different locations on page where search term match were deserved. Title of page, location of keyword, density of keyword, outbound links are important terms to get better ranking position in search engine result [15]. Top factors to improve ranking of results are keyword in URL, keyword in domain name, keyword in H1, keyword in title and density of title tag [14]. Anchor texts terms typically occur in queries which are more likely to be repeated in content [23]. Content based features such as title, meta-description, heading, anchor text, position of keyword and others provide good description about information found on that page. Link based features are taken because search engines like Google rely on incoming and outgoing links in a page. There are some features which were not caught from page content like keyword in domain name, keyword in URL path. Extracted important features are shown in Table 2.

D. Score Computation

After selecting the features precise values have been defined for each feature according to its frequency of occurrence. Of course, there is no general rule to define precise value. By studying research papers and performing some experiments, values are defined to each feature. When search keyword match with the feature found in page then according to criteria value is assigned to feature 'i' called 'fi' and weight to feature 'i' called 'wi'. When all the feature values computed then the total score for a page is computed by adding scores of all features. Flowchart in Fig. 1 shows how total_score is computed for each page.

Table1: Ranking Factors Found In Manual Study

Sr. No.	Ranking factor	Sr. No.	Ranking factor
1	Title tag	27	Links within
2	Meta description tag	28	Outgoing links
3	Meta keyword tag	29	Total links
4	Keyword position in title	30	Keyword in meta title
5	Keyword position in meta_desc	31	Keyword in meta headline
6	Keyword position in meta_keyword	32	Density of keyword in meta_keyword tag
7	Density of keyword in title tag	33	Density of keyword in meta_desc tag
8	Keyword in em tag	34	keyword in div title
9	Keyword in anchor alt	35	keyword in div alt
10	Keyword in anchor text	36	keyword in TD
11	Keyword in A-title	37	Keyword in span
12	Keyword in alt text	38	Keyword in strong
13	Keyword in div	39	Keyword in Bold
14	Keyword in paragraph	40	Keyword in font
15	Keyword in H1	41	Keyword in span
16	Keyword in H2	42	Keyword in strong
17	Keyword in H3	43	Keyword in Bold
18	Keyword in H4	44	Keyword in Big
19	Keyword in H5	45	Keyword in option
20	Keyword in H6	46	Keyword in cufon text
21	Keyword in link title	47	Keyword in cufon alt
22	Keyword in label	48	Keyword in option
23	Keyword in Remark	49	Google hints
24	Keyword in img title	50	Facebook
25	Keyword in area	51	Twitter
26	Keyword in area alt	52	Ads by Google

Table 2: Optimal Subset of Page Features

Sr. No.	Feature
1	Keyword(s) present in title tag
2	Keyword(s) present in meta-description tag
3	Keyword(s) present in domain name
4	Keyword(s) present in URL file path
5	Keyword(s) present in H1 tag
6	Position of keyword in title tag
7	Density of keyword in title tag
8	Keyword(s) present in Anchor text
9	Number of links in a page
10	Number of out/going links in a page

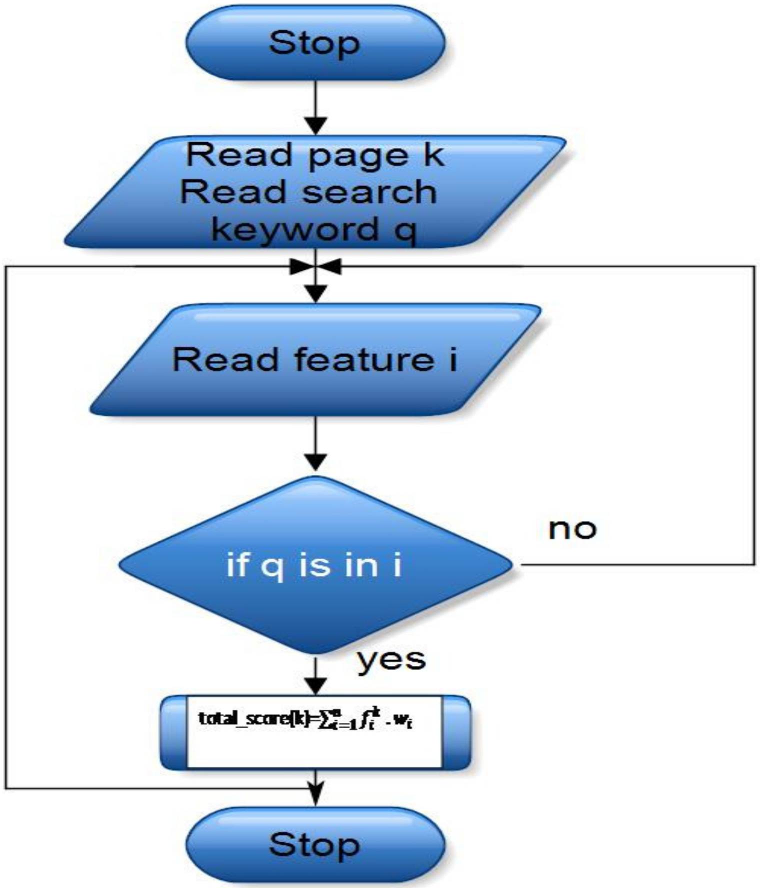

Figure 1: Flowchart for score computation

According to algorithm designed by us reads source code of page, checks feature present or absent and accordingly scores are computed for each feature in a page. Then total score is computed for each page by adding all feature score. This procedure is repeated for all 20 search results. Once all scores are calculated, the web pages are arranged by descending order of total score of a page. This assigned a predicted ranking to each page starting from 1. The criteria assigning precise weights to each feature is shown in table III and the algorithm is shown in fig. 2.

Table 3: Defining Weight to Each Feature

Sr. No.	**Feature**	**Weight W_i**
1	Search keyword present once in Title or meta description or H1 tag	10
	Search keyword present two times in Title, meta description or H1 tag	9
	Search keyword present more than two times in Title tag or meta description tag or H1 tag	0
2	Search keyword in URL path	10
3	Search keyword in Domain name	10
4	Search keyword at first Position in title tag	10
	Search keyword at second Position in title tag	9
	Search keyword at Position greater than second in title tag	0
5	Count in anchor text less than or equal to mean count	10
	Count in anchor text greater than mean count	0
6	Density of search keyword less than or equal to Mean density of title tag	1 to 10
	Density of search keyword greater than Mean density of title tag	-1 to -10
7	Total number of links less than or equal to average of Total links	1 to 10
	Total number of links greater than average of Total links	-1 to -10
8	Total number of outgoing links less than or equal to average of outgoing links	1 to 10

Table 4: Search result by Google, Query 3D TV

Ran k	Site	R/ NR
1	http://en.wikipedia.org/wiki/3D_televisio	Y
2	http://en.wikipedia.org/wiki/WOWvx	N
3	http://en.wikipedia.org/wiki/File:3D_TV.JPG	Y
4	http://3-dtv.org/	Y
5	http://www.3dtv.at/Index_en.aspx	Y
6	http://www.thinkdigit.com/Features/How-3D-TV-works-Part-I-_3568.html	Y
7	http://www.thinkdigit.com/	N
8	http://www.3dtv.at/movies/Index_en.aspx	Y
9	http://www.3dmagic.com/	Y
10	http://www.indiantelevision.com/headlines/y2k10/june/june175.php	Y
11	http://www.indiantelevision.com/headlines/y2k10/june/june80.php	Y
12	http://economictimes.indiatimes.com/articleshow/5660635.cms	Y
13	http://timesofindia.indiatimes.com/articleshow/5825165.cms	Y
14	http://www.3dtv-research.org/	Y
15	/http://www.topnews.in/world-cup-3d-tv-sony-2264623	Y
16	http://timesofindia.indiatimes.com/articleshow/5825165.cms	Y
17	http://www.topnews.in/3d-tv-blamed-pregnancy-2261584	Y
18	http://www.dlp.com/hdtv/dlp-features/3d-hdtv.aspx	Y
19	http://www.honeytechblog.com/samsung-3d-tv-lineup-review/	Y

4. RESULT AND DISCUSSION

Subsequently, predicted ranking and real position are compared for all test pages. It is found that undeserving sites go down while deserving sites get promoted in predicted search results. As a sample, the original results returned for query “3d tv” on Google search engine are in shown in Table IV. In table IV, column heading “Rank” represents ranking position in search result, “Site” represents the name of website returned by Google, “Y” indicates site is relevant to given query and “N” indicates site is irrelevant to given query. Sample snapshot for query “3d tv” is shown in Fig. 2. Table V shows predicted results with total score of a page. According to observation of original and predicted results, irrelevant sites to override and then the quality of search results is improved and now top 10 results contained relevant sites. In Table IV, original results site 2 and site 7 are not relevant to given query but get top ranking position. In predicted results the irrelevant sites computed score becomes low and they move down. Site 2 moves at rank 18 and site 7 moves at rank 20 to bottom of results as shown in Table V.

All the positions of irrelevant sites have been summarized which occurred in original results of Google for predefined queries and position of these sites after executing the algorithm. Table VI shows the comparative evaluation. Comparative evaluation shows irrelevant results move down after re-ranking. The performance of original top 10 search results of Google search engine and the same results after processing and re-ranking for all queries is shown in Table VI. From table VI, it is observed that For query “beauty“, “3d tv” , “deluxe room”, “PSP, the performance increased up to 20%. The performance increased up to 10% for results of query “graphic design” and “iPod” and “pasta pizza”. The original results of queries “House plan” and “insurance” already contained all relevant results in first page search results. For query “Search Engine Optimization” the first page result list contained all relevant sites but the irrelevant site appeared in original result at position 13 moves down to position 20 in predicted results. The predicted results of all queries except for query “beauty” contained all relevant results in first search engine result page.

```
Algorithm Score_compute()
Parameters:
k : web page
q : search keyword
i : feature
n : total number of features
f_i = value presence or absence of feature i in page k]
w_i = weight of feature i
Procedure
Step 1 : Repeat step 2 to 6 for all pages
Step 2 : Read k, q, n
Step 3 : Repeat step 4 to 6 for n features
Step 4 : Read feature i
Step 5 : Check whether search_keyword q matches with
         Keyword in feature i in given page k
Step 6 : If match is found then
        Assign values to f_i and w_i according to given criteria
            $\text{total\_score}(k) = \sum_{i=1}^{n} f_i^k \cdot w_i$
```

Figure 2: algorithm to compute total score of a page

Table 5: Predicted results retured by system

Rank	**Site**	**Total Score**	**R/ NR**
1	http://en.wikipedia.org/wiki/3D_television	80	Y
3	http://en.wikipedia.org/wiki/File:3D_TVJPG	76	Y
18	http://www.dlp.com/hdtv/dlp-features/3d-hdtv.aspx	74	Y
12	http://economictimes.indiatimes.com/articleshow/5660635.cms	72	Y
5	http://www.3dtv.at/Index_en.aspx	60	Y
4	http://3-dtv.org/	56	Y
13	http://timesofindia.indiatimes.com/articleshow/5825165.cms	52	Y
9	http://www.3dmagic.com/	50	Y
15	http://www.topnews.in/world-cup-3d-tv-sony-2264623	48	Y
17	http://www.topnews.in/3d-tv-blamed-pregnancy-2261584	48	Y
16	http://timesofindia.indiatimes.com/articleshow/5825165.cms	47	Y
10	http://www.indiantelevision.com/headlines/y2k10/june/june175.php	41	Y
14	http://www.3dtv-research.org/	40	Y
19	http://www.honeytechblog.com/samung-3d-tv-lineup-review/	36	Y
8	http://www.3dtv.at/movies/Index_en.aspx	30	Y
11	http://www.indiantelevision.com/headlines/y2k10/june/june80.php	29	Y
20	http://www.moneycontrol.com/	27	Y
2	http://en.wikipedia.org/wiki/WOWvx	26	N
6	http://www.thinkdigit.com/Features/How-3D-TV-works-Part-I-_3568.html	20	Y
7	http://www.thinkdigit.com/	10	N

Table 6: Optimal Subset of Page Features

Query	Original Results		Results after Re-ranking	
	R	IR	R	IR
Beauty	70%	30%	90%	10%
3d tv	80%	20%	100%	-
deluxe room	80%	20%	100%	-
graphic design	90%	10%	100%	-
House Plan	100%	-	100%	-
Insurance	100%	-	100%	-
IPod	90%	10%	100%	-
Pasta Pizza	90%	10%	100%	-
PSP	80%	20%	100%	-
Search Engine Optimization	100%	0%	100%	-
Average	**88%**	**12%**	**99%**	**1%**

5. CONCLUSION

Engine optimizers mislead search engines and improved the ranking of pages higher than they deserved. In this paper an algorithm has been proposed to nullify the effect of undeserving sites from search results. As the first step, the investigation of different important features was used by major search engines for ranking. From observations of web search results, their HTML source pages and literature survey were understood which features are commonly used by search engines for ranking. The reason behind this is that these features are most likely the ones that search engine optimizers used. After manual analysis it was found that out of 10 queries the results of 2 queries contained all top sites relevant to given query. After defining precise values and executing algorithm, from results of total 8 queries, predicted results for 7 queries contained all relevant sites in top 10 results and only results for query “beauty” contained 10% irrelevant result in top 10 search list. Average relevancy of all predicted results for predefined queries is improved from 88% to 99%. The system nullifies the effect of undeserving sites from top results. This system effectively and efficiently improves the relevance ranking of web search results in top list.

REFERENCES

[1] Page, L. and Brin, S. (1998). Anatomy of Large Scale Hypertextual Web Search Engine. *Computer Networks*. Vol. 30(1-7):107-117.

[2] Seymour, T., Frantsrog D. and Kumar, S. (2011). History of Search Engines. *Management and Information Systems*. Vol. 15(4):47-58.

[3] Egele, M., Kruegel, K. and Kirda, E. (2011). Removing Web Spam links from Search Engine result. *Journal in Computer Virology, ACM*. Vol.7(1):51-62.

[4] Jansen, B. J. and Spink, A. (2003). Analysis of Web Documents Retrieved and Viewed. *In Proc. 4th Int. Internet Computing*. 65-69.

[5] http://searchenginewatch.com/sew/study/ 2276184/no-1-position-in-google-gets-33-of-search-traffic-study.

[6] Fetterly, D., Manasse, M., and Najork, M. (2004). Spam, Damn Spam and Statistics: Using statistical analysis to locate spam web pages. *In Proc. 7th Int. Web and Databases*. 1-6.

[7] Knezevic, B. and Vidas-Bubanja, M. (2010) Search Engine Marketing As Key Factor for Generating Quality Online Visitors. *In Proc. 33rd Int. Convention*. 1193-1196.

[8] Nursel, Y. and Utku, K. (2010). What is Search Engine Optimization:SEO?, Procedia – Social and Behavirol Sciences, *ELSEVIER*. Vol. 9:487-493.

[9] Lewandowski, D. Web Searching, Search Engines and Information Retrieval. (2005). *Information Services & Use*. Vol. 25(3):137-147.

[10] Henzinger, M. R., Motwani, R. and Silverstein, C. Challenges in web search engines. (2002). *ACM SIGIR Forum.* Vol. 36(2):11-22.

[11] Patil, S. P., Pawar, B. V. and Patil, A. S. (2013). Search Engine Optimization : A Study. *Information Technology and Sciences*. Vol. 1(1):10-13.

[12] Yunfeng, M. A Study on Tactics for Corporate Website Development Aiming at Search Engine Optimization(2010). *IEEE Trans. Computers*. Vol. 3: 673-675.

[13] Cui, M. and Hu, S. (2011). Search Engine Optimization Research for Website Promotion. *IEEE Trans. Computers*. Vol. 4:100-103.

[14] Zhu, C. and Wu, G. (2011). Research and Analysis of Search Engine Optimization Factors Based on Reverse Engineering. *IEEE Trans. Computers*. 225-228.

[15] Wang, F., Li, Y. and Zhang, Y. (2011). An empirical study on the search engine optimization technique and it's outcomes. *IEEE Trans. Computers*. 2767 – 2770.

[16] Fiefei, X. and Guangnian, Z. (2009). Design and Implementation of a Java-based search engine algorithm analysis system. *IEEE Trans. Computers*. 1040-1043.

[17] Shi, J., Cao, Y. and Zhao, X. (2010). Research on SEO Strategies of University Journal Websites. *IEEE Trans. Computers*. 3060-3063.

[18] Su, A., Hu, Y.C., Kuzmanovic, A. and Koh, C. (2010). How to Improve Google Ranking : Myths and Reality. *IEEE Trans. Computers*. 50-57.

[19] Somani, A. and Suman, U. (2011). Counter measure against evolving search engine spamming techniques. *IEEE Trans. Computers*. 214-217.

[20] Spink, A. and Jansen, B. J. (2004). WebSearch : Public Searching of the Web. *Series: Information Science and Knowledge Mgt.* Dordrecht: Kluwer Academic Publishers. Vol. 6:1-199.

[21] Singhal, A. (2005). Challenges in Running a Commercial search Engine: *Proceedings of the 28th annual international ACM SIGIR conference, ACM,* 432-432.

[22] Eiron N. and McCurley K. (2003). "Analysis of Anchor Text for Web Search", *ACM SIGIR*. 459-460.