International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 22 (2015) pp 43298-43300

Performance Evaluation of Search Engine using VSM & User feedback sessions

Siddharth Ghansela[1]

Dept. of Computer Science & Application, G B Pant Engineering College Pauri Garhwal-246194, India

Dr. Ashish Negi[2]

Dept. of Computer Science & Application, G B Pant Engineering College Pauri Garhwal-246194, India

Abstract:
Search engines are developed to help users to rapidly find applicable material on the Internet. A lot numbers of search engine are available in today's searching scenario like Google, Bing, Yahoo, MSN etc. It is observed that most of people use Google for searching any query or document available on the web because of its productive searching breakthrough and rapid connection of database available where as others are struggling with performance issues and tried to search out for better solutions but still satisfactory target is not achieved. Many different techniques for evaluation of search engines performance have been proposed by various researchers. In this paper, we have presented a comparative analysis about the performance of search engine using vector space model, user click through method and Google PageRank checker.

Keywords: Vector, Rank, document, Crawler**.**

1. Background

From the previous 13 years search engines plays an important role in information retrieval. The first searching tool named Archie was created by Alan Emtage in 1990 [1]. After that Gopher was introduced by Mark Mccahill in 1991 [2]. A web crawler was developed by Matthew Gray at MIT in 1993 [3]. Another search engine, Aliweb also comes in 1993 [4]. The first crawler-based search engine was introduced in 1994 by web [5]. Around 2001, the Google search engine rose to prominence. In 2004, Yahoo launched its own search engine. In 2005, MSN by Microsoft launch its search engine.

2. Analytical Model

An analytical model represents record and dubiety usually as angle, forge or directions. The similarity of the questioning direction is represented as a scald cost. There are two types of layouts available namely first magnitude (numerical basis) and second magnitude (properties of the prototypical) [6-8]. In brief several analytical layouts are available today are listed as below.

a. Vector Space Model
b. Postulate VSM
c. Upgrade VSM
d. Enlarged Boolean Model
e. Latent Semantic Indexing

In this paper we have implemented the vector space model because it allows computing a continuous degree of similarity between queries and documents, and it is easy to implement.

3. Vector Space Model

Vector space prototype or term angle layout is an algebraic model for representing content documents [6]. It is used in data retrieval, indexing and evaluation of documents. Its first use was in SMART data retrieval system [6]. In the vector space model, we represent data as angle. Statistical model, that consisting Vector Space prototype [7-9] and Probabilistic layout helped much and became the criterion for their architecture and algorithms.

The achievement or downfall of the vector space method is based on phrase density [10]. Phrase density is a technique by which we can calculate the weight of a term in the given number of documents. The classical method for computation of phrase density [11] is given by $Wij = tf * idf$. tf is term frequency (number of terms) and idf is inverse document frequency (global information). idf is scaling factor, it tells the important of terms. If query terms appear in many documents, and its importance will be scaled down. The idf [12] is given by $idf = \log \frac{D}{df_j}$, where D is the number of record in the record store and $tf_{Q,j} \times idf_j$ number of documents containing the query term. If $df_j < D$, the term will have large Idf value. The similarity function [13] between documents vectors D and query Q is given by,

$$Cos\theta = Sim(Q, D_i) = \frac{\sum_{j=1}^{V} W_{Q,j} \times W_{i,j}}{\sqrt{\sum_{j=1}^{V} W_{Q,j}^2} \times \sqrt{\sum_{j=1}^{V} W_{i,j}^2}}$$

where, $W_{Q,j}$ is the weight of term j in the query Q, and is defined similar way as $W_{i,j}$ (i.e., $tf_{Q,j} \times idf_j$.)

4. Google Page Rank Algorithm

Google page rank method was given by Sergey Brin and Lawrence Page in 1998[14]. This design advise us the approach for devious the rank of a page among the set of

record. Further it is also applicable to calculate the rank of a page. Google page rank algorithm works on crawling and indexing. It first collects all the links of a website (crawling) and then arranges them into the database (indexing). Google page rank algorithm is given by the following formula:

PR(A) = (1-d) + d (PR(T1)/C(T1) +... + PR(Tn)/C(Tn))

Where

PR(A) = Page Rank of Page *A*

PR(T1) - PR(Tn) –pages correlate to page "*A*".

PR(T1) …PR(Tn) – the first and last pages which bond to "*A*" as well as every web page in middle.

PR(Tn)/C(Tn) – any linking web page divides the weight of its poll constantly between all of the polls that it gives.

d – damping factor (commonly.85).

Now at present scenario Google search engine works over 200 factors.

5. Click Through Methods

Click division is a feature recommended by Lee [15]. Generally, a period for web exploration is an array of subsequent queries to appease a single information need and some clicked search results. The proposed feedback session is based on clicked URLs. The single period includes all the three URLs. Each feedback session can tell what a user wants and what he/she does not requires. Therefore, for inferring user search goals, it is more efficient to analyse the feedback sessions than to analyse the search results or clicked URLs directly.

6. Proposed Approach

In this paper we have implemented the VSM and click through method by taking a query and then compare the results from the above mentioned methods by the Google Page rank checker. Here we have taken Google search engine for searching a query, and from the out coming results we have taken first three documents from the first page of search engine. Then from those result we calculate the rank of all the three documents with respect to the given query by using Vector Space Model. After calculating the result using Vector Space Model we then compare the rank with Google page rank auditor. In the last section we will compare the out coming results.

7. Experimental Results

The query selected is as follows:

Q: *Government Engineering College in Uttarakhand* and the results are given as follows:

D1: Govind Ballabh Pant Engineering colleges Pauri Garhwal Uttarakhand

D2: Top Government Engineering colleges in Uttarakhand

D3: Top engineering colleges in Uttarakhand 2013

On applying the VSM on the above results.

Terms	Term in Q	Count tf_i			df	$\frac{D}{df_i}$	$\log(\frac{D}{df_i})$ IDF_i	Weights, $W_i = tf_i \times IDF_i$			
		D1	D2	D3				Q	D1	D2	D3
In	1	0	1	1	2	1.5	0.1761	0.1761	0.0000	0.1761	0.1761
Government	1	0	1	0	1	3.0	0.4772	0.4772	0.0000	0.4772	0.0000
Engineering	1	1	1	1	3	1.0	0.0000	0.0000	0.0000	0.0000	0.0000
College	1	1	1	1	3	1.0	0.0000	0.0000	0.0000	0.0000	0.0000
Uttarakhand	1	1	1	1	3	1.0	0.0000	0.0000	0.0000	0.0000	0.0000
Govind	0	1	0	0	1	3.0	0.4772	0.0000	0.4772	0.0000	0.0000
Ballabh	0	1	0	0	1	3.0	0.4772	0.0000	0.4772	0.0000	0.0000
Pant	0	1	0	0	1	3.0	0.4772	0.0000	0.4772	0.0000	0.0000
Pauri	0	1	0	0	1	3.0	0.4772	0.0000	0.4772	0.0000	0.0000
Garhwal	0	1	0	0	1	3.0	0.4772	0.0000	0.4772	0.0000	0.0000
Top	0	0	1	1	2	1.5	0.1761	0.0000	0.0000	0.1761	0.1761
2013	0	0	0	1	1	3.0	0.4772	0.0000	0.0000	0.0000	0.4772

8. Similarity Analysis

The similarity function is

$|D_1|$=1.0671

$|D_2|$=1.0082

$|D_3|$=0.5383

$|Q| = 1.1432$

$Q.D_1 = 0$

$Q.D_2 = 1.3066$

$Q.D_3 = 0.3522$

$$cos\, Q.D_1 = \frac{Q.D_1}{|Q| \times |D_1|}$$

So from Vector prototype method the rank of all the three documents are given as follows:

$Cos\, Q.D1 = 0.0000$

$Cos\, Q.D2 = 1.3337$

$Cos\, Q.D3 = 1.0000$

9. Comparison of methods

After comparing the three methods the results and rank comparison of the three documents are shown below.

Table-1.1 Rank of documents based on three methods

Document	Rank Based On		
	Vector Space Model	Google Page Rank Checker	User Click Through (Points out of 10)
D1	0.0000	5	8
D2	1.3337	3	6
D3	1.0000	1	7

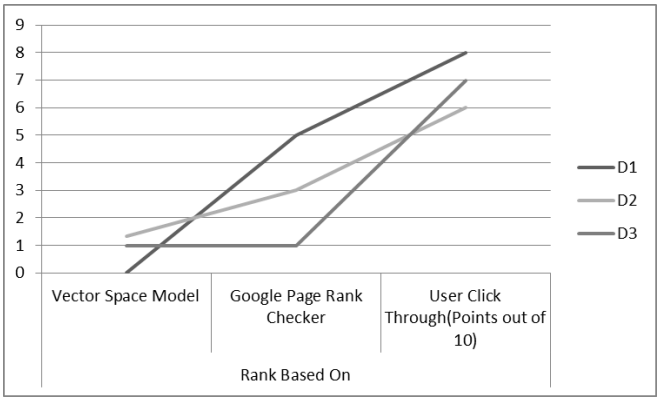

Fig. 1.2 Rank comparison of three methods

10. Conclusion and future work

In this paper we have taken one query based on the user thought not based on TREC standard and the three out coming results are shown above. We have three results for each document, for document (D1) we get rank 3 for VSM, rank 1 for Google Page Rank Checker and user click through method, for document (D2) we get rank 1 for VSM, rank 2 for Google Page Rank Checker and rank 3 for user click through method and at last for document (D3) we get rank 2 for VSM, rank 3 for Google Page Rank Checker and rank 2 for user click through method. So we concluded that if we add user feedback as a step in search engine prototype or algorithm we can get better results. In future we can apply different queries in various search engines that can give more strength in our results.

11. References

[1] Leiden, U., Archie, 2001, "Internet History - Search Engines" (from Search Engine Watch), Universite it Leiden, Netherlands, web: Leiden U-Archie.

[2] Randall, J., and Neil, 1994, "The World Wide Web unleashed" Sams Publishing. pp. 20.

[3] Gray, M., Internet Growth and Statistics: Credits and background. http://www.mit.edu/people/mkgray/net/background.html.

[4] Koster, M., 1993, "ANNOUNCEMENT: ALIWEB (Archie-Like Indexing for the WEB)". comp.infosystems (plaintext version).

[5] Croft, B., Metzler, D., Strohman, T., 2010, "Search Engines: Information Retrieval in Practice" ISBN-10: 0136072240, Pearson.

[6] Shalton, G., Wong, A., and Yang, C., S., 1975, "A vector space Model for automatic indexing" Communications of The ACM, Volume 18, Issue 11.

[7] Dwivedi, S., K., Singh, J., N., and Gotam, R., 2011, "Information Retrieval Evaluative Model" FTICT: Proceedings of the 2011, International conference on "Future Trend in Information & Communication Technology, Ghaziabad, India.

[8] Longzhuang, L., and Shang, Y., 2000, "A new statistical method for performance evaluation of search engines".

[9] Longzhuang, L., and Shang, Y., 2000, "A new method for automatic performance comparison of search engines".

[10] Polettini, N., 2004, "The Vector Space Model in Information Retrieval - Term Weighting Problem".

[11] Buckley, C., 1993, "The importance of proper weighting methods", Proceedings of the Workshop on Human Language Technology, (HLT '93), ACM Press, Stroudsburg, PA, USA., pp. 349-352.

[12] Salton, G., and C. Buckley, 1988, "Term-weighting approaches in automatic text retrieval", pp. 513-523.

[13] Lee, D., L., Chuang, H., and Seamons, K., 1997, "Document ranking and the vector-space model" IEEE Trans. Software, 14: 67-75. DOI: 10.1109/52.582976.

[14] Brin, S., and Page L., 1998, "The anatomy of large-scale hyper textual web search engine", pp. 107-117.

[15] Lee, U., Liu, Z., and Cho, J., "Automatic Identification of User Goals in Web Search" 14th WWW Conference, 2005.