International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 379-383 Jan 2018-June 2018 www.csjournals.com

Performance Analysis of Search Engines based on Similarity Score of Web Pages

Suprity, Jaswinder Singh
[1]M.Tech Scholar, [2]Assistant Professor, Department of Computer Science & Engineering,
Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India

Abstract**:** *The performance of three search engines Google, Bing and AltaVista is analyzed in the paper. Primarily queries were entered in each engine then using the top ten documents were selected on the basis of the top ten links and m words were selected. Then all selected keywords are combined to form single keyword and are transformed into binary form. After that the fitness value is considered and Dice similarity measure was used as fitness function and genetic operators are applied. Now again the new keywords are added in to the queries and the whole process is repeated again. The new similarity value is obtained and then compared with the old value. On the basis of similarity score, the performances of the search engines have been analyzed in this paper.*
Keywords: Search Engine, Relevancy, Similarity measure, Genetic Algorithm.

I. Introduction

Search engine has become a useful tool for people to find pertinent information from the World Wide Web. But some problems are coupled with search engine such as superfluous information, out-dated information and inappropriate information. WWW is based on hypertext and is ever-increasing as a comprehensive information system day by day. The rapid extend of WWW results in the disappointment of search engines to search the newest information. Search System is an IRS that searches for sites rooted in the terms that are nominated as query. Search engines gaze from beginning to end their individual database of information so as to uncover what content is that user are searching for. When user enters request in the form of query then the matching method of the search system delivers the ranked list of documents to the user using the similarity measures. The database containing pages, query system and matching method are three fundamental components of IRS [1], [2], [3]. If the user is not fulfilled with the results returned by search system then user reformulates query there by increasing the retrieval effectiveness iteratively and incrementally [2]. The user evaluates the results on the basis of retrieved documents and provides the relevant feedback for the expansion of terms of initial query. This paper contains five sections. The first section of explains the introduction about search engine. The second section of paper describes the work related to the selection of similarity measure. The third section of paper describes the methodology & experimentation and fourth section of paper describes the results. The fifth section of paper describes the conclusion.

II. Related work

Many efforts have been done by the various researchers to develop efficient system to retrieve the relevant documents but it was difficult for the system to retrieve the documents when only two or three terms are added in the search box of the system to retrieve the relevant documents. So there is need to explore the methods related to the query enhancement or expansion and to design the similarity function for the effective information retrieval as well as for increasing the visibility of the search. The documents retrieved from the web are in the different forms but the major content is the text and the similarity of the text can computed with the string similarity functions. The string based similarity functions were further classified as the term based similarity functions and character based similarity function. Jaccard, Cosine, Dice and Overlap similarity functions are called as term base similarity coefficient. It was concluded from the literature that the discussed similarity functions were placed in the identical class by most of the authors as described in [4], [5], [6], [7], [8], [9], [11], [12]. Dice similarity between the set of terms of first document set i.e. X and set of terms of second document set i.e. Y is defined as follows.

$$D(X,Y) = \frac{2|X \cap Y|}{|X|+|Y|}$$

III. Methodology & Experimentation

Three search engines i.e. Google, Bing and AltaVista have been analyzed on the basis of their relevancies using dice function. MATLAB was used in the experiments. First of all the queries were entered in the search engine,

the first ten links related to queries were taken and from these links, the text of the related documents were analysed using the Textalyser tool [10] and calculated its frequency on the basis of which the key words were selected. Now these words are combined to form keyword set and converted into binary chromosomes [13]. The fitness value is calculated using the dice coefficient similarity function. Genetic operators like selection, crossover and mutation are applied and now the best keyword was selected. The selected keyword is added with query which was entered initially, now again the new chromosomes are built and the whole procedure is repeated again. The new fitness value is obtained now and this one is compared with the old fitness value. Following are the tables showing the values and the comparisons.

IV. Results

The similarity score of the retrieved documents using Google, Bing and AltaVista search engine using Dice similarity is shown in table1, table 2 and table 3 respectively.

Old Keyword	New Added Term	Best Value With Old Keyword	Best Value With added new Keyword
Indian economy	Government	0.4926	0.5130
2G scam	Minister	0.4191	0.4330
Terrorism in India	Pakistan	0.4413	0.5540
Corruption in India	Lokpal	0.4654	0.7468
Jan Lokpal bill	Corruption	0.5831	0.6333
Indian Railway system	Rail	0.5044	0.5491
World Health Organization	International	0.4774	0.5147
IT sector in India	Growth	0.5547	0.5604
Indian education system	Schools	0.5055	0.5697
Search Engine Optimization	Internet	0.4702	0.4912

Table1: Similarity Score for Google

Old Keyword	New Added Term	Best Values With Old Keyword	Best Value With added new Keyword
Indian economy	Business	0.5678	0.5922
2G scam	India	0.5864	0.5957
Terrorism in India	Attacks	0.5802	0.5841
Corruption in India	Government	0.6077	0.6610
Jan Lokpal bill	Corruption	0.5941	0.6278
Indian Railway system	Information	0.5339	0.5426
World Health Organization	State	0.5466	0.5922
IT sector in India	Companies	0.5411	0.5417
Indian education system	Students	0.6072	0.6607
Search Engine Optimization	Keyword	0.5452	0.5861

Table1: Similarity Score for Bing

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 379-383 Jan 2018-June 2018 www.csjournals.com

Old Keyword	New Added Term	Best Value With Old Keyword	Best Value With added new Keyword
Indian economy	Global	0.4596	0.6103
2G scam	Court	0.4342	0.4349
Terrorism in India	Border	0.4155	0.4600
Corruption in India	Movement	0.5879	0.6120
Jan Lokpal bill	Anna	0.6058	0.6301
Indian Railway system	Services	0.5665	0.6278
World Health Organization	Social	0.4916	0.5722
IT sector in India	Software	0.4899	0.5968
Indian education system	Government	0.5378	0.5414
Search Engine Optimization	Google	0.5619	0.6136

Table3: Similarity Score for AltaVista

The graphs were drawn for the above said results and are shown in fig.1, fig. 2 and fig. 3 respectively. In the graphs queries are represented on X axis and the similarity is represented on Y axis. From the figures it is clear that the similarity score have been increased when new terms are added to the terms of original query.

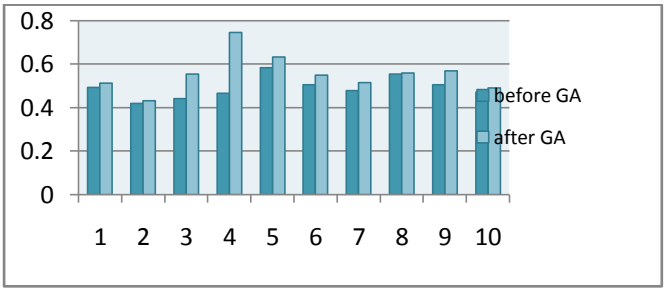

Fig. 1 Similarity Score for Google

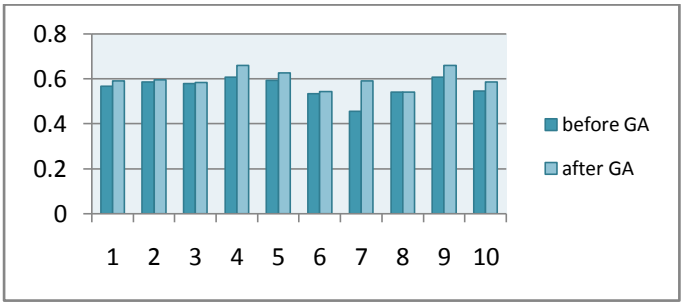

Fig. 2 Similarity Score for Bing

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 379-383 Jan 2018-June 2018 www.csjournals.com

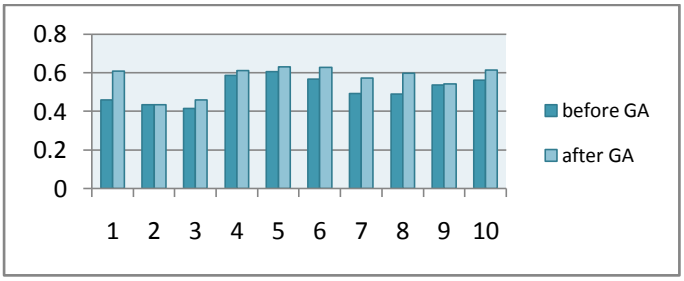

Fig. 3 Similarity Score for AltaVista

The percentage improvement in similarity score is calculated after addition of new terms in the original query for each search engine i.e. Google, Bing and AltaVista and is shown in table 4.

For Search engine	Percentage Similarity improvement Google	Percentage Similarity improvement Bing	Percentage Similarity improvement Altvista
query1	2	2.4	15.07
query2	1.39	.93	0.07
query3	11.27	.39	4.45
query4	28.14	5.3	2.41
query5	5.02	3.7	2.43
query6	4.4	.87	6.1
query7	3.7	4.5	8.06
query8	0.5	.06	10.69
query9	6.4	5.35	0.36
query10	2.1	4.09	5.17

Table 4 : Percentage Improvement in Similarity Score

V. Conclusion

In this paper, a procedure is described to compare three search engines by using similarity measure dice coefficient and genetic algorithm. The similarity score of the search engines have been improved with the addition of new terms in the original query terms. Further, the percentage improvement in old and new fitness values for Google comes out to be best then AltaVista and then Bing.

References

[1] R. Baeza-Yates and B. Ribiero-Neto, *Modern Information Retrieval*. Addison Wesley, New York, 1999.

[2] V. N. Gudivada, V. V. Raghavan, W. I. Grosky, and R. Kasanagottu, "Information retrieval on the world wide web," *IEEE Internet Computing*, no. 5, pp. 58–68, 1997.

[3] Michael Gordon, "Probabilistic and genetic algorithms in document retrieval," *Communications of ACM*, vol.31, no. 10, pages. 1208-1218, 1988.

[4] M. Bilenko and R. J. Mooney, "Adaptive duplicate detection using learnable string similarity measures," *Proc. 9th ACM SIGKDD, Int. Conf. Knowledge Discovery and Data Mining, KDD-2003*, Washington DC, USA, 2003 pp. 39-48.

[5] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, 2013.

[6] L. Egghe and C. Michel, "Strong similarity measures for ordered sets of documents in information retrieval," *Information Processing & Management*, vol. 38, no. 6, pp. 823–848, 2002.

[7] T. P. vander Weide and P. van Bommel, "Measuring the incremental information value of documents," *Information Sciences*, vol. 176, no. 2, pp. 91–119, 2006.

[8] Sung-Hyuk Cha, "Comprehensive survey on the distant/similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, Issue 4, pp. 300-307, 2007.

[9] M.-C. Kim and K.-S. Choi, "A comparison of collocation-based similarity measures in query expansion," *Information Processing & Management*, vol. 35, no. 1, pp. 19–30, 1999.

[10] *http://textalyser.net*.

[11] Jaswinder Singh, Parvinder Singh, Yogesh Chaba," A study of Similarity Functions Used in Textual Information Retrieval in Wide Area Networks", *International Journal of Computer Science & Information Technologies*, vol. 5, No.6, pp. 7880-7884, 2014.

[12] Jaswinder Singh, " Search Term Expansion using Dice Similarity Measure" *International Journal of Electronics Engineering*, vol.9, issue 2, pp. 308-314, 2017.

[13] Z. Michalewicz, *Genetic Algorithm + Data structure = Evolution programs*. Springer, 1996.