International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 927-931 June - Dec 2018 www.csjournals.com

Performance Modeling of Information Retrieval System based on Relevancy of Retrieved Documents

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

Abstract: Today, Web is key resource of information. With the increasing use of internet, It is expanding in size and complexity day by day. But the main challenge of the web is to provide the relevant documents effectively and efficiently. In this paper performance of Google, MSN and AltaVista search engines are compared on the basis of relevancy of pages that are retrieved from each search engine. Relevancy of page is calculated by Jaccard coefficient similarity measure in MATLAB environment.Textalyser tool analyses the pages retrieved from search engine and returns list of important terms. These terns are converted into binary form. Fitness is then calculated using Jaccard similarity Coefficient and genetic operators are applied which results into a new term. This new term is then added to original query.Fitness value is again calculated and result is obtained.
Keywords: Search Engine, Relevancy, Similarity measure, Genetic Algorithm.

I. Introduction

Search engine is used as tool to search the information regarding his/ her topic of interest. Users enter the terms or phrases and terms related to their choice of interest then the search engine generates a list of relevant web pages. Retrieval of documents is based on the extent of the resemblance between terms entered by the user and the pages provided by Search Engines. This means that a document with more resemblance value to the entered term is judged to be more significant to it and should be provided by the retrieval system in a superior location in the catalog of pages provided by search engine [1].Today a very large number of search engines are there which are used as tool for searching the information e.g. Google, Yahoo!, MSN, Alta Vista, Bing. The similarity measurement between the different objects is the fundamental function of any information retrieval application and there are varieties of ways to compute the similarity among the different object representations. In the field of information retrieval, the two objects may be the two different documents or one document and query. The documents retrieved from the web are in the different forms but the major content is the text. Text of the documents contains the words, sentences and paragraphs and most of the content on the web is in English language. Similarity of the text can computed with the string similarity measures. The effectiveness of IRS system is described in first section of paper. The second section of paper describes the literature related to similarity measure and the methodology details. Third, fourth and fifth section describes the experimental details, results and analysis respectively. The conclusion is described in sixth section of paper.

II. Related Work & Methodology

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 in addition to matching method are three primary components of Retrieval System [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 [1]. The user evaluates the results on the basis of retrieved documents and provides the relevant feedback for the expansion of terms of initial query. In this paper, only text is considered for studying its impact on the accessibility of search system. Queries were chosen for retrieving the web pages from the web by using search engine. In the experiment ten queries were chosen which are described in the table1. Additionally, three information retrieval system were chosen i.e. Google, MSN and ALTAVSTA.

Table 1: Queries used in experiment

Query No.	Query
Q1	Corruption in India
Q2	Indian Education System
Q3	IT Sector in India
Q4	World Health Organization
Q5	Indian Railway System

Q6	Jan Lokpal Bill
Q7	Search Engine Optimization
Q8	Terrorism in India
Q9	2G Scam
Q10	Indian Economy

In the information retrieval system the similarity as well as the relevancy of the retrieved pages relies on the similarity measures. Therefore the importance of the results relies on selection of similarity measures. In this paper Jaccard similarity coefficient was used. For X and Y subsets of documents retrieved from the entire repository of documents. The formula for Jaccard similarity function was defined in [1], [2], [4], [5], [6], [8], [9]. Jaccard similarity between the set of terms of first document set i.e. X and the set of terms of second document set i.e. Y is defined as follows.

$$J(X,Y) = \frac{|X \cap Y|}{|X|+|Y|-|X \cap Y|}$$

III. Experimentation

In this work similarity measure is used as a fitness function and Genetic Algorithm is applied. Results from fitness function are interval 0 to 1. This experimentation is done on ten queries in search engines Google**,** MSN and AltaVista using Jaccard similarity measure as fitness function. When search engine receives the query term, it gives results in form of documents. These documents are sorted in terms of relevancy. After considering the documents, terms are extracted from documents by Textalyser tool [7]. Chromosomes are encoded in form of binary [10]. All terms of documents are arranged in ascending order in form of a set and these chromosomes are termed as the initial population that is fed into genetic operators. The terms of a set which is present in a document is assigned one otherwise zero. Relevancy of documents for search engine is calculated by Jaccard coefficient fitness function. After evaluating population's fitness, selection operator is applied. Selection operator selects only those chromosomes which have higher fitness value. Here roulette wheel selection is used for this purpose. After this cross over and mutation operator are applied. In this experiment, Jaccard function gives best convergence of relevancy value at (Crossover probability=0.8 and Mutation probability=0.01). Next, optimized query chromosome is decoded into query. The keyword which is repeated in most documents; present in the query chromosome and which is highly relevant as compared to other keywords is added to the original query. Only one keyword is added to the original query. Average relevancy with new keyword is calculated.

IV. Results

The relevancy of the retrieved documents using Google, MSN and AltaVista search engine using Jaccard similarity is shown in table1, table 2 and table 3 respectively.

Query No.	Query Entered in Search Engine(Google)	Relevancy of documents with original query	New Added term	Relevancy of documents with added term
Q1	Corruption in India	0.3387	Government	0.4575
Q2	Indian Education System	0.3848	Schools	0.5414
Q3	IT Sector in India	0.4105	Industry	0.4334
Q4	World Health Organization	0.3556	Global	0.4292
Q5	Indian Railway System	0.3830	Trains	0.3887
Q6	Jan Lokpal Bill	0.4462	Corruption	0.5691
Q7	Search Engine Optimization	0.3413	Google	0.4123
Q8	Terrorism in India	0.3176	Punishment	0.3559
Q9	2G Scam	0.3097	Telecom	0.4485
Q10	Indian Economy	0.3674	Bank	0.4019

Table 1: Relevancy of retrieved documents using Google Search Engine

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 927-931 June - Dec 2018 www.csjournals.com

Query No.	Query Entered in Search Engine(MSN)	Relevancy of documents with original query	New Added term	Relevancy of documents with added term
Q1	Corruption in India	0.4731	Political	0.4820
Q2	Indian Education System	0.4770	Student	0.4987
Q3	IT Sector in India	0.4069	Business	0.4425
Q4	World Health Organization	0.4037	Children	0.4256
Q5	Indian Railway System	0.4144	Budget	0.4545
Q6	Jan Lokpal Bill	0.4664	Corruption	0.4752
Q7	Search Engine Optimization	0.3990	Google	0.4929
Q8	Terrorism in India	0.4413	Attacks	0.4691
Q9	2G Scam	0.4409	Telecom	0.4889
Q10	Indian Economy	0.4345	Growth	0.5188

Table 2: Relevancy of retrieved documents using MSN Search Engine

Query No.	Query Entered in Search Engine(Alta Vista)	Relevancy of documents with original query	New Added term	Relevancy of documents with added term
Q1	Corruption in India	0.4517	Money	0.4995
Q2	Indian Education System	0.4100	Government	0.4685
Q3	IT Sector in India	0.3655	Industry	0.3766
Q4	World Health Organization	0.3595	Global	0.4292
Q5	Indian Railway System	0.4368	Train	0.4537
Q6	Jan Lokpal Bill	0.4767	Anna	0.4943
Q7	Search Engine Optimization	0.4164	Marketing	0.5130
Q8	Terrorism in India	0.2983	Mumbai	0.4021
Q9	2G Scam	0.3149	Spectrum	0.5079
Q10	Indian Economy	0.3357	Government	0.4055

Table 3: Relevancy of retrieved documents using Alta Vista Search Engine

Fig.1 shows the graph between relevancy (Y axis) and queries (X axis) for the Google, MSN and AltaVista respectively for the original search terms and fig. 2 shows the graph for relevancy with new terms for the said search engines.

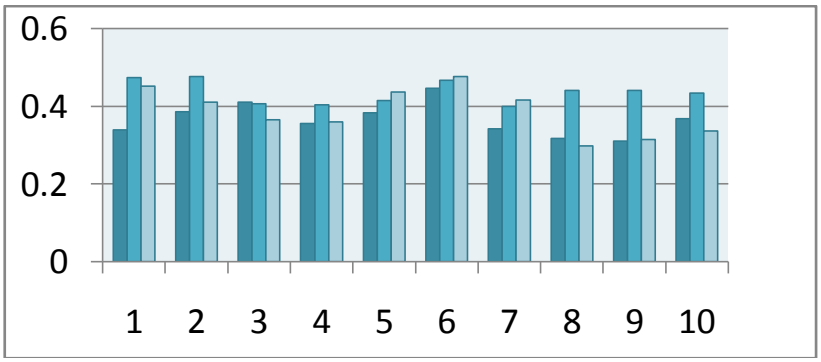

Fig. 1Graph of Old Terms of Google, MSN and AltaVista

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 927-931 June - Dec 2018 www.csjournals.com

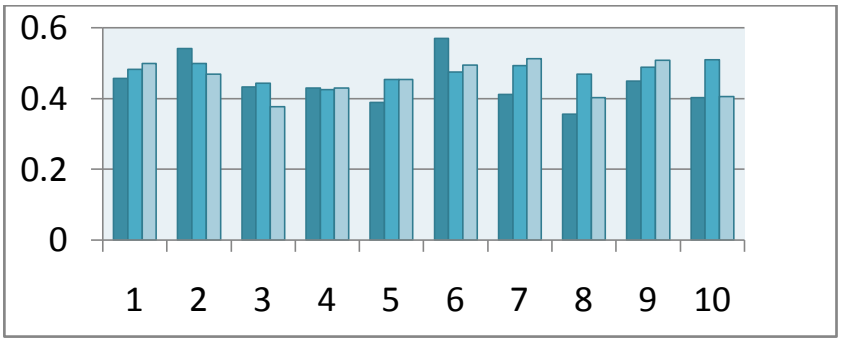

Fig. 2 Graph of New Terms of Google, MSN, Alta Vista

The comparison of relevancy of retrieved documents for the old search terms and new terms is shown in table 4 and the percentage improvement in relevancy for Google, MSN and AltaVista is shown in table 5.

Query No.	Relevancy of documents with original query (Google)	Relevancy of documents with new query (Google)	Relevancy of documents with original query (MSN)	Relevancy of documents with new query (MSN)	Relevancy of documents with original query (ALTA VISTA)	Relevancy of documents with new query (ALTA VISTA)
Q1	0.3387	0.4575	0.4731	0.4820	0.4517	0.4995
Q2	0.3858	0.5414	0.4770	0.4987	0.4100	0.4685
Q3	0.4105	0.4334	0.4069	0.4425	0.3655	0.3766
Q4	0.3556	0.4292	0.4037	0.4256	0.3595	0.4292
Q5	0.3830	0.3887	0.4144	0.4545	0.4368	0.4537
Q6	0.4462	0.5691	0.4664	0.4752	0.4767	0.4943
Q7	0.3413	0.4123	0.3990	0.4929	0.4164	0.5130
Q8	0.3176	0.3559	0.4413	0.4691	0.2983	0.4021
Q9	0.3097	0.4485	0.4409	0.4889	0.3149	0.5079
Q10	0.3674	0.4019	0.4345	0.5088	0.3357	0.4055

Table 4: Comparison of Relevancy of Old Terms and New Terms in Google, MSN and AltaVista

Query No.	Google	MSN	ALTA VISTA
Q1	11.880	0.89	4.78
Q2	15.560	02.170	05.850
Q3	04.760	03.560	01.110
Q4	13.640	02.190	06.970
Q5	0.570	04.010	01.690
Q6	12.29	0.880	01.780
Q7	7.09	9.39	9.66
Q8	03.83	02.780	10.380
Q9	13.88	04.800	19.300
Q10	03.45	07.430	6.980

Table 5: Percentage Improvement in Relevancy of Google, MSN and AltaVista

Search Engine	Average Percentage Improvement in Relevancy
Google	8.695
MSN	3.810
Alta Vista	6.832

Table 6: Average Percentage Improvement in Relevancy of Google, MSN and AltaVista

V. Result Analysis

Experiment is performed on the same set of ten queries on Google, MSN, and AltaVista search engines. As shown from the graph average relevancy of new keywords is high as compared to old keywords. Fig. 1 and fig. 2 show the graph of old and new keywords for Google, MSN and AltaVista before applying genetic algorithm and after applying genetic algorithms respectively. Table 6 shows the Average Percentage improvement for Google, MSN and AltaVista respectively. Result shows that Google gives maximum percentage improvement in comparison with other search engines; MSN have least percentage increase in relevancy and average percentage increase in relevancy of AltaVista lies between Google and MSN.

VI. Conclusion

So it is concluded that performance of retrieval systems (Google, MSN and AltaVista) which was analyzed on the basis of relevancy of retrieved documents is best in Google and percentage increase in relevancy of retrieved documents of AltaVista lies between Google and MSN.

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] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, New York, USA, 1986.

[5] Jaswinder Singh, Parvinder Singh, Yogesh Chaba,“ Performance Modelling of Information Retrieval Techniques Using Similarity Functions in Wide Area Networks,” International Journal of Advanced Research in Computer Science and Software Engineering, vol.4, issue 12, pp.786-793, 2014.

[6] 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.

[7] http://textalyser.net.

[8] 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.

[9] Jaswinder Singh, “ Expanding Query using Jaccard Similarity Measure”, International Journal of Computer Science & Communication, vol. 8, issue 1,pp.70-75, 2017.

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