International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 372-378 Jan 2018-June 2018 www.csjournals.com

Expansion of Search Terms using Overlap Similarity Measure

Jaswinder Singh
Assistant Professor, Department of Computer Science & Engineering,
Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India

Abstract:
When user enters request in the form of query then the ranked list of documents is delivered by the matching mechanism of the search system. But the user usually enters two or three words as query so it difficult for the user to get the relevant information.These short queries with two or three terms and the incompatibility among terms of search query and the pages affect the relevancy of retrieved pages. In this paper Genetic Algorithm is applied and Overlap similarity measure is used as fitness function. With the implementation of Genetic Algorithm, new terms are accessible and then similarity is measured by adding new term in the old terms and percentage increase in the similarity is measured.

Keywords: Overlap Similarity Measure, Genetic Algorithm, Search Term expansion.

I. Introduction

The widespread text retrieval task from web globe is to recover pages in reply to the user's search. The information required is retrieved using the information access process. This process is frequently used by searcher while searching the content from the web globe. The information need of the user is different for different users and their behavior during search is also different during search. On the basis of search strategies, searchers are generally classified into three different categories i.e. an informal user, an investigator looking for serious content and a specialized user [1]. The information need of the user may be categorized as the informative, navigational and transactional. To search any information the user from the web, a search system is often chosen by the user and then query is formulated by the user according to the information need and as per the requirement of the search system. On the basis of information need, the queries can be classified as the informative search terms, navigational terms and transactional terms. The query is then processed by search system and then the search system returns the pages related to terms of query. These documents are then evaluated by user that whether these are related to his or her information requirement. Significance of document is personal in nature as it rely on the opinion of the user [2].The goal of the search component of the search system is to forecast about the pages that are applicable to the user and position the pages in the predicted possibility of the significance of user. Pages with additional resemblance with request of search terms entered by searcher will have more relevancies and will be at superior spot in the listing of the returned pages. The relevancy can be considered by measuring resemblance among pages retrieved and search terms by via the similarity measures [2]. If the need of the user is fulfilled then this process of information access and retrieval stops otherwise the user reformulate the query by enhancing the query by adding some more keywords or reformulate the query by completely changing the keywords because the retrieval process and information access process is iterative search process. The database containing pages, query system and matching method are three fundamental components of IRS [2], [3], [4]. 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 [3]. The user evaluates the results on the basis of retrieved documents and provides the relevant feedback for the expansion of terms of initial search. Query expansion is a technique used to increase the effectiveness of the information retrieval [2]. It is the process of adding some more terms or phrases to the existing query to improve relevancy of the retrieved documents. The reformulated query contains more terms so the probability of matching them with terms in relevant documents is therefore enhanced. The first section of explains the introduction about retrieval system. The second section of paper describes the work related to similarity measure and expansion of query. The third section describes the methodology followed and about the experimentation. The fourth section of paper describes the results and conclusion is described in section five of paper.

II. Related Work

An introduction to information access process and IRS described in the preceding section and is shown in fig.1. Many efforts have been done by the various researchers to develop such a system but it was difficult for the system to retrieve the relevant documents when only two or three terms are added in the search box of the system. So there is need to explore the methods related to the enhancement of searched terms and to design the similarity function for the effective information retrieval as well as for increasing the accessibility of the search. When user enters the search term to search the topic of his/her interest, the search system returns the documents with appropriate links. 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.It was found that the string based similarity functions were further classified as the term based similarity functions and character based similarity function. Jaccard [17], Cosine [18], Dice [19] and Overlap similarity functions are called as term base similarity coefficient [6]. From literature it is concluded that above said similarity functions were placed in the identical class by most of the authors as described in [5], [6], [7], [8], [9], [10], [11], [12], [20] and vector space model is applicable to all these similarity functions. In general the character based similarity functions are used to compare the short strings and it is because of this reason it is too expensive to apply them for the large documents so the token base similarity functions or term-based similarity functions avoids these problems by viewing as the bag of terms or tokens.

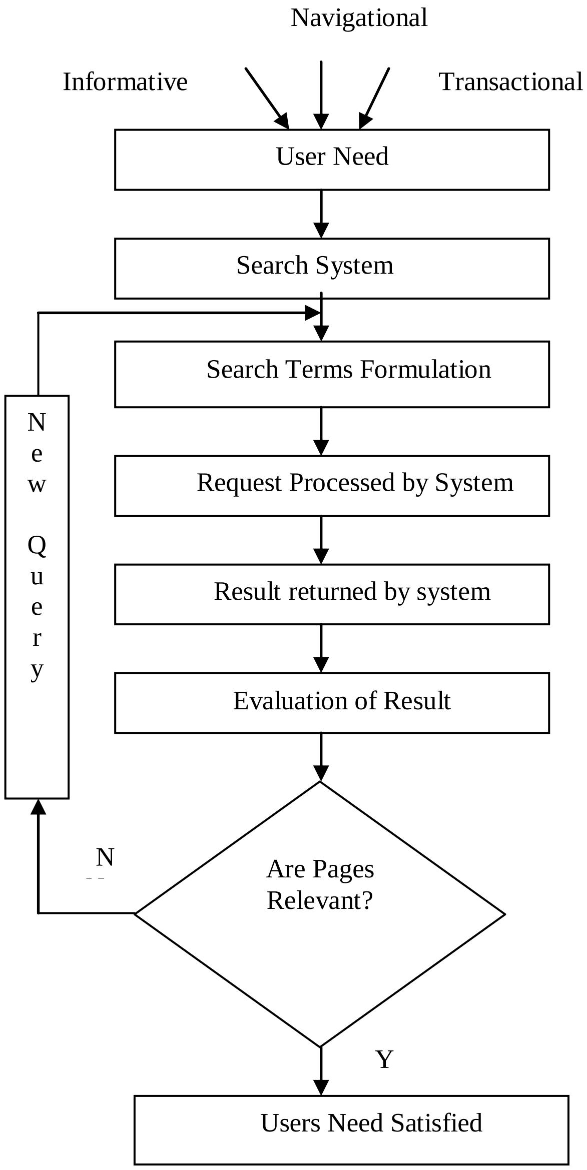


Fig. 1: Information Access Process.

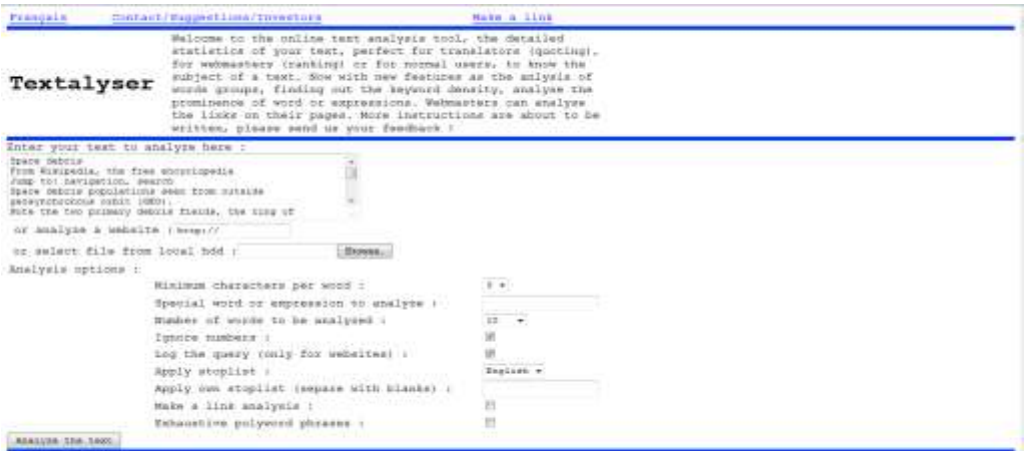
From the literature related to the expansion techniques of search terms, it was found that the relevancy of the retrieved documents can be improved by gathering of more suitable terms with query terms as with the addition of additional terms in query there is more likelihood of retrieving the correct documents. It was found from the literature that out of local analysis and global analysis, local analysis is enhanced method to inflate the search terms in which the ranked documents are retrieved first and then important words from the pages are extracted and then added to the terms of query this will enhance relevancy of the retrieved documents. By studying the literature it was also found that the genetic algorithm is the good method for the optimization. The implementation of Genetic Algorithm and methodology followed to achieve the results is described in the next section of paper.

III. Methodology & Experimentation followed

In this paper, only text is considered for studying its impact on the accessibility of search system. The methodology and experimentation includes following steps.

Step1: Preparation of Data

The experimentation starts with the preparation of data. Queries were chosen for retrieving the web pages from the web by using Google search engine. In the experiment ten queries were chosen. Additionally, it is also required to choose the structure of data for the experimentation. Query Q1is "Terrorist Attack Mumbai".
In response to query Q1, after retrieving top ten documents, keywords were collected from documents using the tool i.e.Textalyser [13].



Welcome to the online text analysis tool, the detailed statistics of your text, perfect for translators (quoting), for webmasters (ranking) or for normal users, to know the subject of a text. Now with new features as the anlysis of words groups, finding out the keyword density, analyse the prominence of word or expressions. Webmasters can analyse the links on their pages. More instructions are about to be written, please send us your feedback !

Enter your text to analyze here :

Space debris
From Wikipedia, the free encyclopedia
Jump to: navigation, search
Space debris populations seen from outside geosynchronous orbit (GEO).
Note the two primary debris fields, the ring of

or analyze a website : http://

or select file from local hdd :

Analysis options :

Minimum characters per word : 3
Special word or expression to analyze :
Number of words to be analyzed : 10
Ignore numbers : ☑
Log the query (only for websites) : ☑
Apply stoplist : English
Apply own stoplist (separe with blanks) :
Make a link analysis : ☐
Exhaustive polyword phrases : ☐

Fig. 2: Snap Shot of Textalyser Tool

Keyword set of 25 terms (in the experiment) i.e. significant terms of all the top ten documents was formed for the search term i.e. "Terrorist Attack Mumbai". Chromosomes are encoded in form of binary [14], [15]. Keywords of documents are arranged in ascending order in form of a set as it was described in [16]. These chromosomes are called initial populations that are fed into genetic operators. The code for fitness evaluation and implementation of genetic algorithm is done using MATLAB. The encoding using binary weights was done for query Q1 as shown below.

```
x =[0,1,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,1,1,0,1,0,1,0;
    0,0,0,0,0,0,0,1,0,0,1,0,0,0,1,1,0,0,0,0,1,0,0,0,1;
    0,1,0,1,0,0,0,0,0,0,1,1,0,0,0,0,1,1,1,0,0,1,0,0,0;
    0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0;
    0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,1,0,0,0,1;
    0,1,0,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,1,0,0,0,1,0,1;
    0,1,0,1,0,0,1,0,1,0,1,0,0,0,0,0,0,0,1,1,0,0,0,0,1;
    1,1,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,1,0,0,0;
    0,1,0,0,0,1,0,0,1,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,1;
    0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,1;
   ];
```

In similar way the data related to the document representation was prepared for the ten selected queries i.e. Q1,Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10 using the above said tool and document representation for the ten selected queries were done.

Step 2: Fitness evaluation using overlap similarity function

After the preparation of data the similarity measure was chosen from the literature. The performance analysis of the similarity functions i.e. Overlap is done with the help of the training data in the experiments. For example, Similarity measure measures the extent of similarity between two sub sets X and Y of entire data base of the documents in the repository i.e.

"X is defined, a set of all terms occurring in document X
Y is set of all terms occurring in document Y.
|X| = Numbers of terms that occur in set X.
|Y| = Number of terms that occur in set Y.
$| X \cap Y |$ =Number of terms occur in both X and Y."

For X and Y subsets of documents retrieved from the entire repository of documents. The formula for the Overlap similarity measure was defined in [7], [2], [12], [16]. Overlap similarity measure is classified as similarity measure based on term [7], [8]. Overlap 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.

$$O(X,Y) = \frac{| X \cap Y |}{min(|X|,|Y|)}$$

In the experiment, number of terms that occur in both i.e. Doc1 and Doc2 are measured and min of (Doc1terms ,Doc2 terms) is measured then by using Overlap similarity function Olp1is calculated as shown below.

Doc1 = 0100000000100010001101010

Doc1 = 0100000000100010001101010

Olp1 = $O(Doc1, Doc1) = \frac{|7|}{\min(|7|,|7|)} = 1$

Table 1: Overlap similarities for search term "Terrorist Attack Mumbai"

Docs	Similarity value using Overlap Similarity measure with search terms										**Avg.**
	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	
Doc1	1	0.3333	0.5714	0.6666	0.6	0.4285	0.5714	0.5	0.5	0.25	0.5421
Doc2	0.3333	1	0.1666	0.3333	0.6	0.3333	0.3333	0	0.3333	0.25	0.3683
Doc3	0.5714	0.1666	1	0.6666	0.6	0.4285	0.5	0.5	0.5	0.25	0.5183
Doc4	0.6666	0.3333	0.6666	1	0.3333	0.3333	0.3333	0.3333	0.3333	0	0.4333
Doc5	0.6	0.6	0.6	0.3333	1	0.8	0.8	0.4	0.8	0.25	0.6183
Doc6	0.4285	0.3333	0.4285	0.3333	0.8	1	0.7142	0.3333	0.8333	0.25	0.5454
Doc7	0.5714	0.3333	0.5	0.3333	0.8	0.7142	1	0.3333	0.8333	0.25	0.5669
Doc8	0.5	0	0.5	0.3333	0.4	0.3333	0.3333	1	0.3333	0.25	0.3983
Doc9	0.5	0.3333	0.5	0.3333	0.8	0.8333	0.8333	0.3333	1	0.25	0.5716
Doc10	0.25	0.25	0.25	0	0.25	0.25	0.25	0.25	0.25	1	0.3

The value of overlap was obtained as "Olp1, Olp2, Olp3, Olp4, Olp5, Olp6, Olp7, Olp8, Olp9 and Olp10" as shown in table1. Then the average of all the coefficients was calculated. The process up to this point is repeated with other chosen queries i.e. Q2,………Q10 and the result is shown in table2.

Table 2: Overlap similarity with entered search terms

Query No.	Search Terms Entered in Search Engine	Relevancy of documents with old search terms
Q1	Terrorist Attack Mumbai	0.4863
Q2	Cloud Burst India	0.3427
Q3	Moist Attack India	0.3960
Q4	Corruption Cricket India	0.4592
Q5	Pollution River Ganga	0.6645
Q6	Power Generation India	0.4269
Q7	Sand Mining India	0.5675
Q8	Mid Day Meal India	0.4949
Q9	Sikh Riots India	0.5141
Q10	Moist Attack Train	0.5627

Step3: Search Term Expansion

After studying the literature it was concluded that the local feedback is good technique for the expansion of search terms and this technique is implemented by applying the genetic algorithm. While applying the genetic algorithm it is required to have the training data in proper form of strings of zeros and ones so as to form the population to apply the genetic algorithm operators. After evaluating population's fitness, the next step is chromosome selection. Selection operator selects only those chromosomes which have higher fitness value. Here roulette wheel selection was used for this purpose. After this cross over and mutation operator were applied. The experiment was done for 500 generations with probability of crossover i.e. 0.5 and mutation probability .001. The experiment was repeated with different probability of crossover and mutation rates. New keyword was chosen by selecting the bit position which has value one in the mutated chromosome. Only one keyword is added to the original query. Average similarity with new keyword is calculated and the results are shown in table 3.

IV. Results

The experimentation of the Overlap similarity function used as the fitness function in the genetic algorithm is performed using the same method as described in section third of this paper. Similarity between the retrieved documents is measured using the Overlap similarity functions for the newly formulated query which is formed by adding the keyword i.e. "Headly Terrorist Attack Mumbai". The method of addition of more terms is described in [17].The Process of relevancy measurement is repeated for new query as described in previous section.

Doc1'= 0,1,1,0,0,0,0,0,1,0,0,1,0,0,1,0,1,1,0,0,1,0,0,0,0;
Doc1'= 0,1,1,0,0,0,0,0,1,0,0,1,0,0,1,0,1,1,0,0,1,0,0,0,0;

$$\text{Olp1'} = O(Doc1', Doc1') = \frac{|8|}{\min(|8|,|8|)} = 1$$

Doc1'= 0,1,1,0,0,0,0,0,1,0,0,1,0,0,1,0,1,1,0,0,1,0,0,0,0;
Doc2'= 0,1,0,0,1,0,0,0,1,0,0,1,0,0,1,0,1,0,1,0,0,0,0,0,0;

$$\text{Olp2'} = O(Doc1', Doc2') = \frac{|5|}{\min(|8|,|7|)} = 0.7142$$

Table 3: Overlap similarities for expanded search terms "Headly Terrorist Attack Mumbai"

Docs	Similarity value using Overlap Similarity measure with new search terms										Avg.
	Doc1'	Doc2'	Doc3'	Doc4'	Doc5'	Doc6'	Doc7'	Doc8'	Doc9'	Doc10'	
Doc1'	1	0.7142	0.5714	0.5714	0.7142	0.6666	0.5714	0.5	0.2857	0.5	0.6095
Doc2'	0.7142	1	0.4285	0.4285	0.5714	0.5	0.4285	0.3333	0.2857	0.5	0.519
Doc3'	0.5714	0.4285	1	0.8571	0.5714	0.5	0.7142	0.6666	0.2857	0.3333	0.5929
Doc4'	0.5714	0.4285	0.8571	1	0.5714	0.5	0.7142	0.6666	0.2857	0.3333	0.5929
Doc5'	0.7142	0.5714	0.5714	0.5714	1	0.6666	0.7142	0.6666	0.2857	0.5	0.6262
Doc6'	0.6666	0.5	0.5	0.5	0.6666	1	0.6666	0.3333	0.3333	0.3333	0.55
Doc7'	0.5714	0.4285	0.7142	0.7142	0.7142	0.6666	1	0.5	0.2857	0.3333	0.5929
Doc8'	0.5	0.3333	0.6666	0.6666	0.6666	0.3333	0.5	1	0.3333	0.5	0.55
Doc9'	0.2857	0.2857	0.2857	0.2857	0.2857	0.3333	0.2857	0.3333	1	0.5	0.3881
Doc10'	0.5	0.5	0.3333	0.3333	0.5	0.3333	0.3333	0.5	0.5	1	0.4833

The similarity value was obtained in the experiment as "Olp3', Olp4', Olp5', Olp6', Olp7', Olp8', Olp9' and Olp10' ". The above process is repeated as described in previous section and is shown in table 3. In the experiment similarity using Overlap similarity function is measured with the expanded search terms i.e. Q1' and average value of similarity is obtained which is 0.5505.This value is compared with the value of previous query i.e. Q1 which is 0.4863 and percentage improvement in similarity is calculated which is 6.42 % in the case of Overlap and this result is shown in the following table 4.

Table 4: Relevancy of retrieved documents for query Q1 with added term

Query No.	Query Entered in Search Engine	Newly Added term	Relevancy of document before adding term	Relevancy of retrieved documents after adding term	Percentage Improvement
Q1	Terrorist Attack Mumbai	Headly	0.4863	0.5505	6.42

The process was repeated with the data with the other queries i.e. Q2, Q3,.. Q10, the results obtained are summarized in the table 5 and the results shows that there is improvement in the relevancy of the retrieved documents when the term returned by genetic algorithm was added into the original query.

Table 5: Relevancy of retrieved documents using Overlap similarity function with expanded term

Query No.	Query Entered in Search Engine	Relevancy of documents with original query	New Added term	Relevancy of documents with added term	Percentage Improvement
Q1	Terrorist Attack Mumbai	0.4863	Headly	0.5505	6.42
Q2	Cloud Burst India	0.3427	Uttarakhand	0.5279	18.52
Q3	Moist Attack India	0.3960	Train	0.5627	16.67
Q4	Corruption Cricket India	0.4592	Fixing	0.5445	8.53
Q5	Pollution River Ganga	0.6645	Industrial	0.6870	2.25
Q6	Power Generation India	0.4269	Thermal	0.5385	11.16
Q7	Sand Mining India	0.5675	Illegal	0.6239	5.64
Q8	Mid Day Meal India	0.4949	Bihar	0.5513	5.64
Q9	Sikh Riots India	0.5141	Sajjan	0.6539	13.98
Q10	Moist Attack Train	0.5627	People	0.6078	4.51

V. Conclusion

The similarity of retrieved documents is improved by using the Overlap similarity measure as the fitness function in Genetic Algorithm and percentage enhance in the similarity is measured. With the spreading out of the terms of search, it was found that additional terms are available for the search system. With the availability of more terms the accessibility of search increases.

References

[1] S. Srinivasa and P. C. P. Bhatt, "Introduction to Web information retrieval: A user perspective," *Resonance*, vol. 7, no. 6, pp. 27–38, 2002.

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

[3] 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.

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

[5] W. P. Jones and G. W. Furnas, "Pictures of relevance: A geometric analysis of similarity measures," *Journal of the American Society for Information Science*, vol. 38, no. 6, pp. 420–442, 1987.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] Jaswinder Singh, Parvinder Singh, Yogesh Chaba," Increasing the visibility of search using Genetic Algorithm," *Journal of Computer Engineering*, vol.17, issue 5,ver.1,pp.7-17, 2015.

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

[16] Jaswinder Singh, Parvinder Singh, Yogesh Chaba," Performance Modeling 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.

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

[18] Jaswinder Singh, " Expanding Search Accessibility using Cosine Similarity Measure" International Journal of IT & Knowledge Management, vol. 10, No.2, pp. 136-140, 2017.

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

[20] Y. Amirgaliyev and B. Bakiyev, "Comparing Text Based Documents Similarity Measuring Functions," *Proc. 12th International Conf. Information Technologies and Management*, Riga 2014.