Volume 4, Issue 12, December 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Performance Modelling of Information Retrieval Techniques Using Similarity Functions in Wide Area Networks

Jaswinder Singh[1], Parvinder Singh[2], Yogesh Chaba[3]
[1,3]Department of Computer Science& Engineering, Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India
[2] Department of Computer Science & Engineering, Deenbandhu Chhotu Ram University of Science &Technology, Murthal, Sonepat, Haryana, India

Abstract— The World Wide Web is the largest repository of public data and it is continuously expanding in size and complexity with the increasing use of internet but to retrieve the relevant documents is still a big challenge in the field of information retrieval in wide area networks. In this paper the focus is on the text and the textual similarity is measured using the different similarity functions which are used as information retrieval techniques in the field of information retrieval in wide area networks. Various similarity functions have been developed but how they are best applied in information retrieval and how similarity values should be interpreted is not answered yet. In this paper the performance modelling of information retrieval techniques using the different similarity functions or similarity measures i.e. Jaccard, Cosine, Dice and Overlap has been done in wide area networks by measuring the similarity between the keywords entered by the user in the search engine and the documents retrieved for the entered keywords. In this paper ten queries were entered for the experiment, the queries and the documents were represented using the Vector space model, Google search engine was used as search tool to retrieve the first ten documents, the Textalyser tool was used to extract the keywords from the retrieved documents and coding for the similarity calculation was done in MATLAB.

Keywords— Performance Modelling, Information Retrieval Techniques, Similarity Functions, Wide Area Networks

I. INTRODUCTION

The web is expanding continuously and the users of internet are also increasing as a world-wide computer network can be accessed via a computer, mobile phones, PDA, digital TV etc. Today around 40% of world population has internet connection. The number of internet users has increased ten fold from 1999 to 2013[1].The web is expanding continuously because of increase in the amount of online text. The request of the different types of information have developed the interest of researchers in the various fields of information and retrieval in wide area networks like multimedia retrieval, text and topic detection ,chemical and biological informatics etc. [2]. But whenever the user put his or her query at any search engine firstly the text is compared so the textual similarity is the fundamental to all the above said fields as most of information retrieval requests are text based and the textual similarity functions plays the crucial role in the text related research and the tasks related to its applications in the field of information retrieval i.e. text classification and topic detection. The tool used to retrieve the relevant information from wide area networks is called search engine and it was claimed by a survey that 85% of internet users use search engines to find specific information of interest [3]. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms [4] and it was found that approximately 15% of queries submitted have never been seen before by Google's search engine on day to day basis [5]. The objective of search engine is to provide relevant results to the user but the locations of documents, data formats, access mechanism, communication protocols etc. are hidden from users [6]. As most of the queries for the retrieval requests are text based so the focus of the paper is on the keyword based search. The textual similarity measures are the similarity functions or similarity coefficients which are used to measure the degree of similarity between user query and documents. When the user enters the query at the search engine site, then the user input is checked against the search index of all the pages it has analysed, the best results are then returned to the user. The goal of this paper is to analyse the performance of the different similarity functions in wide area networks. The remainder of paper is organized as follows.

The first section of paper describes the brief introduction about the increase in usage of internet and about the key role of textual similarity while retrieving the information. The second section describes the information retrieval techniques in wide area networks with brief introduction of information retrieval system. Third section describes the various similarity functions which are used frequently in the field of information retrieval in wide area networks.The fourth section discusses the tools which were used in the experiment. Fifth section of paper describes the detailed explanation about the experiment for the similarity measurement between query and documents using the Jaccard, Cosine, Dice and Overlap similarity functions. Sixth section of paper is about the analysis and conclusions.

II. INFORMATION RETRIEVAL TECHNIQUES IN WIDE AREA NETWORKS

The basic component of any information system is the representation of the information itself. In the textual information retrieval, representation means the representation of documents and queries. Representation of queries means the representation of user request. Nicholas J. Belkin et.al [7] described that a technique used for comparing the query and document is called the retrieval technique and classified the retrieval techniques as the exact match techniques and partial match techniques. The exact match techniques excludes the relevant texts whose representations match the query only partially but the partial match techniques are those in which the retrieved documents will also include those documents that exactly match with the query. Nicholas J.Belkin et.al [7] further classified the partial match techniques as the individual document representation and the techniques that use a representation of documents that stresses on the connections to the other documents in the network. Individual techniques further classified as the structure-based and the feature-based where the documents are represented by the sets of features and the index terms. The vector space model is based upon the formal model of document retrieval and indexing is the key representative of the feature based category. This model designates texts and queries as vectors in a multidimensional space, the dimensions of which are the words used to represent texts. Queries and texts are then compared by comparing the vectors as in the case of cosine similarity coefficient [8] [9]. The vector space model is based on the assumption that more similar a vector representing a text is to a query vector, the more is likely that the text is relevant to that query [6]. An information retrieval system is defined as a system in which the contents of information items are interpreted and ranking is generated which reflects relevance and retrieves the information more efficiently. Most of Information retrieval system uses the keywords to retrieve the documents. The systems first extract keywords from documents and then assign weights to the keywords by using different approaches. Information retrieval system consists of three basic components i.e. Documentary Database, Query Subsystem and Matching mechanism [9] [10]. The locations of documents and data formats etc. are usually hidden from users and this document database stores document along with the information content of their representation [6]. Query subsystem is a system which formulate user request into query. Matching mechanism compares the similarity between the query and documents in the database. Based on this, documents are retrieved. The utility of the information retrieval system depends on the performance of the similarity functions and the performance of the similarity function further depends on the semantic sensitivity to different kinds of that may be contained in the database matrix of document and term weights [11].

III. RELATED WORK

It was found from the literature that wide range of similarity functions have been developed which are used in the different fields such as information retrieval [12], image retrieval [13], molecular ecology [14], genetics and molecular biology [15] and chemistry [16]. McGill et.al [17] surveyed the different similarity measures. Sung-Hyuk Cha [12] described the different categorization of similarity measures for comparing the nominal type of histograms. William P. Jones et.al [11] described the geometric representation of similarity measures i.e. Inner Product, Cosine, Dice and Overlap using the vector space model. Wael H. Gomaa et.al [18] characterized the textual similarity functions into String-based, Corpus-based and knowledge-based. String based approach is further categorised as the character-based approach and the term-based approach .The term-based approach includes Jaccard, Cosine, Dice and Overlap similarity functions. Suphakit Niwattanakul et.al [19] concluded that Jaccard similarity coefficient is suitable sufficiently to be employed in the word similarity measurement. Wa'el Musa Hadi et.al [20] concluded that the Cosine outperforms the Dice and Jaccard similarity as different variations of vector space model were compared. Xiaojun Wan [21] proposed the novel similarity measure based upon earth movers distance. In the information retrieval, similarity functions are functions which are used to measure the degree of similarity between user query and documents. The simplest way of counting the documents and query is by counting the number of terms they have common. Retrieving documents in response to a user query is the most common text retrieval task. For this reason, most of the text similarity functions have been developed that take input as a query and retrieve the matching documents. Various similarity functions have been developed but how they are best applied in information retrieval and how similarity values or rankings should be interpreted is not answered yet because the main problem with the similarity measure is that each measure is related to particular application domain. It is therefore difficult to decide which similarity function should be used for a particular application and which information model is used. From the literature [11], [20], [21] it was found that the vector space model can be applied to the Jaccard, Cosine, Dice and Overlap similarity functions. These similarity functions fall in the same category of term-based similarity measure as described by Wael H. Gomaa et.al [18] after completing the survey on categorization of the text based similarity approaches and these also belong to the same family as described by Sung-Hyuk Cha [12] in the comprehensive survey of similarity measures.

In this paper four similarity functions have been considered for performance analysis using the vector space model in wide area networks i.e. Jaccard, Cosine, Dice and Overlap. Binary weights have been used and weight of term is considered as '1' if term occurs in the document and weight of term is considered as '0' if the term does not occur in the document.

X is defined, a set of all terms occurring in document X

Y is set of all terms occurring in document Y.

$|\mathbf{X}|$ = Numbers of terms that occur in set X.

$|\mathbf{Y}|$ = Number of terms that occur in set Y.

$|\mathbf{X} \cap \mathbf{Y}|$ = Number of terms occur in both X and Y.

IV. RESEARCH TOOLS

A. Textalyser Tool:

It is a powerful online tool which gives statistics about a text including word count, unique words, number of sentences, average words per sentence and lexical density [22].This tool was used in the experiment to extract the top keywords from top ten documents retrieved by the Google search engine.

B. Search Engine:

In the experiment Google search engine was used as the search tool to search the relevant pages for the given query.

C. MATLAB:

MATLAB is high-level language and interactive environment [23]. It was used in the experiment for similarity measurement i.e. Jaccard similarity, Cosine similarity, Dice similarity and Overlap similarity.

V. EXPERIMENTATION

In the experiment Google search engine was used for retrieving the documents and ten documents were retrieved for one query. The experiment was done for the ten queries and four similarity functions were used for measuring the similarity between the document and query. Process of experiment done is as follows.

Step1. Information request i.e. query was entered in search box of search engine.
Step2. Keywords were extracted from the retrieved documents to form the keyword set for query.
Step3. Documents were represented for query and weights were assigned.
Step4. Textual Similarity of retrieved web pages for entered query using similarity function was measured.
Step5. The above steps were repeated with different similarity functions and with different queries.
Step6. Textual Similarity was compared.
Step7. Performance analysis of different similarity functions was done.

A. Extraction of keywords for the query "Terrorist Attack Mumbai" to form the set of keywords

After the retrieval of top ten documents, keywords were extracted from documents using the Textalyser tool.
When user enter query into search engine, it gives results in form of documents which are already sorted in terms of relevancy i.e. document with higher relevancy gets first priority over less relevant document as returned by the search engine. In the experiment when Query= **{Terrorist Attack Mumbai}** was typed in the search box of Google search engine then top ten documents displayed by Google search engine were considered and following keywords were extracted using the Textalyser tool . In the experiment the first ten documents were represented as D1, D2, D3, D4, D5, D6, D7, D8, D9and D10.The text of these documents were considered for the extraction of keywords related to the query.

D1 = {Attack, Mumbai, Pakistan, India, Police, Kasab, Taj}
D2 = {India, Terrorist, People, Government, killed}
D3 = {Mumbai, Blast, Police, Attack, India, Maharashtra, Injured, Minister}
D4 = {Kasab, Mumbai, Attack}
D5 = {India, Terrorist, Mumbai, Pakistan, Attack}
D6 = {Headly, India, Rana, Attack, Hillary Clinton, Terror}
D7 = {Mumbai, Headly, Attack, Terror, Friday, India, Pakistan, Blast}
D8 = {Police, Mumbai, Alert, Attack, Bandra, Juhu, Afzal}
D9 = {Mumbai, Attack, Headly, Case, Terror, India},
D10= {Intelligence, Police, Terror, Bomb}

The following keyword set of 25 terms representing all the ten documents was formed for the entered query:
{Afzal, Attack, Bandra, Blast, Bomb, Case, Friday, Government, Headly, Hillary Clinton, India, Injured, Intelligence, Juhu, Kasab, Killed, Maharashtra, Minister, Mumbai, Pakistan, People, Police, Rana, Taj, Terrorist}

B. Document Representation for the query "Terrorist Attack Mumbai"

The term "Afzal" of the keyword set was compared with the terms of document, D1 which do not contain the term "Afzal" so '0' was placed at first position for the document, D1. Similarly second term "Attack," of keyword set was compared with the terms of document, D1 which contain the term "Attack" so '1' was placed at second position of the document, D1. In this way the encoding using binary weights was done to obtain the ten documents i.e. D1, D2, D3, D4, D5, D6, D7, D8, D9 and D10 of length '25' as there are 25 keywords in the extracted keyword set for the entered query.

D1 = 0100000000100010001101010;
D2 = 0000000100100011000010001;
D3 = 0101000000110000111001000;
D4 = 0100000000000010010000000;
D5 = 0100000000100000001010001;
D6 = 0100000011100000001000101;

D7 = 0101001010100000001100001;
D8 = 1110000000000100001001000;
D9 = 0100010010100000001000001;
D10= 0000100000001000000001001;

C. Jaccard Similarity, Cosine Similarity, Dice Similarity & Overlap Similarity for query "Terrorist Attack Mumbai"

In the experiment Jaccard similarity, Cosine similarity, Dice similarity and Overlap similarity was measured for query.

1. Jaccard Similarity:

Jaccard Similarity function was used as first similarity function in the experiment is also known as Jaccard coefficient or Jaccard index and it is member of family of term-based similarity measure [18]. For text document, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms. In the experiment the terms of first document, D1 were compared with terms of first document, D1 using Jaccard similarity function and value of Jaccard coefficient is obtained as Jac1which is '1' as shown in table1.In the same way the terms of first document, D1 were compared with the terms of second document, D2 using the Jaccard similarity function to obtain the value of Jaccard coefficient as Jac2. In the similar manner Jaccard coefficient were obtained as Jac1, Jac2, Jac3, Jac4, Jac5, Jac6, Jac7, Jac8, Jac9 and Jac10. Then the average of all the coefficients was obtained. The above process was repeated for all other documents and the average of all the coefficients was taken as shown in table 1 and following formula of Jaccard coefficient was used in the experiment for the Jaccard similarity.

$$\frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}$$

TABLE1. JACCARD SIMILARITY FOR QUERY "TERRORIST ATTACK MUMBAI"

Doc	**Similarity with Jaccard Similarity Function**										**Avg.**
	Jac1	**Jac2**	**Jac3**	**Jac4**	**Jac5**	**Jac6**	**Jac7**	**Jac8**	**Jac9**	**Jac10**	
D1	1	0.1818	0.3636	0.25	0.3333	0.2727	0.3636	0.3	0.3	0.1	0.3465
D2	0.1818	1	0.0769	0.125	0.375	0.1818	0.1666	0	0.2	0.1111	0.2418
D3	0.3636	0.07692	1	0.2222	0.3	0.25	0.3333	0.2727	0.2727	0.0909	0.3182
D4	0.25	0.125	0.2222	1	0.1428	0.1111	0.1	0.125	0.125	0	0.2201
D5	0.3333	0.375	0.3	0.1428	1	0.5	0.4444	0.2222	0.5714	0.125	0.4014
D6	0.2727	0.1818	0.25	0.1111	0.5	1	0.5	0.1818	0.625	0.1	0.3722
D7	0.3636	0.1666	0.3333	0.1	0.4444	0.5	1	0.1666	0.5555	0.0909	0.3721
D8	0.3	0	0.2727	0.125	0.2222	0.1818	0.1666	1	0.2	0.1111	0.2579
D9	0.3	0.2	0.2727	0.125	0.5714	0.625	0.5555	0.2	1	0.1111	0.3960
D10	0.1	0.1111	0.0909	0	0.125	0.1	0.0909	0.1111	0.1111	1	0.1840

2. Cosine Similarity:

Cosine Similarity function [8] [9] was used as second similarity function in the experiment. If Cosine value is one then two documents are identical and zero if nothing is common between them. It also falls in the category of term-based similarity measure [18].In the experiment the first document terms were compared with terms of first document using Cosine similarity function to obtain the value of Cosine similarity as Cos1 and the terms of first document, D1 were compared with terms of second document, D2 using the Cosine similarity function to obtain the value as Cos2 and so on. The values of Cosine were obtained as Cos1, Cos2, Cos3, Cos4, Cos5, Cos6, Cos7, Cos8, Cos9 and Cos10. The above process was repeated for all other documents and the average of all the coefficients is taken as shown in table 2 and following formula of Cosine coefficient was used for the Cosine similarity.

$$\frac{|X \cap Y|}{|X|^{1/2} . |Y|^{1/2}}$$

TABLE2. COSINE SIMILARITY FOR QUERY "TERRORIST ATTACK MUMBAI"

Doc	Similarity with Cosine Similarity Function										Avg.
	Cos1	Cos2	Cos3	Cos4	Cos5	Cos6	Cos7	Cos8	Cos9	Cos10	
D1	1	0.3086	0.5345	0.4364	0.5070	0.4285	0.5345	0.4629	0.4629	0.1889	0.4864
D2	0.3086	1	0.1443	0.2357	0.5477	0.3086	0.2886	0	0.3333	0.2041	0.3371
D3	0.5345	0.1443	1	0.4082	0.4743	0.4008	0.5	0.4330	0.4330	0.17677	0.4505
D4	0.4364	0.2357	0.4082	1	0.2581	0.2182	0.2041	0.2357	0.2357	0	0.3232
D5	0.5070	0.5477	0.4743	0.2581	1	0.6761	0.6324	0.3651	0.7302	0.2236	0.5414
D6	0.4285	0.3086	0.4008	0.2182	0.6761	1	0.6681	0.3086	0.7715	0.18898	0.4969
D7	0.5345	0.2886	0.5	0.2041	0.6324	0.6681	1	0.2886	0.7216	0.1767	0.5015
D8	0.4629	0	0.4330	0.2357	0.3651	0.3086	0.2886	1	0.3333 3	0.20412	0.3631
D9	0.4629	0.3333	0.4330	0.2357	0.7302	0.7715	0.7216	0.3333	1	0.20412	0.5225
D10	0.1889	0.2041	0.1767	0	0.2236	0.1889	0.17677	0.2041	0.2041	1	0.2567

3. Dice Similarity:

Dice Similarity function was used as third similarity function in the experiment and is also known as Dice coefficient. Dice coefficient is defined as twice the number of common terms in the compared strings divided by the total number of terms in the both strings [18]. It is also a member of term-based similarity measure [18]. In the experiment the first document terms were compared with terms of first document using Dice similarity function to obtain the value as Dic1 and the terms of first document, D1 were compared with the terms of second document, D2 using the dice similarity function to obtain the value as Dic2 and so on. The values of Dice similarity were obtained as Dic1, Dic2, Dic3, Dic4, Dic5, Dic6, Dic7, Dic8, Dic9 and Dic10. Then the average of all the coefficients is obtained. The above process was repeated for all other documents and the average of all the coefficients is taken as shown in table 3.The Dice's score formulation is given below:

$$\frac{2\,|X \cap Y|}{|X| + |Y|}$$

TABLE3. DICE SIMILARITY FOR QUERY "TERRORIST ATTACK MUMBAI"

Doc	Similarity with Dice Similarity Function										Avg.
	Dic1	Dic2	Dic3	Dic4	Dic5	Dic6	Dic7	Dic8	Dic9	Dic10	
D1	1	0.3076	0.5333	0.4	0.5	0.4285	0.5333	0.4615	0.4615	0.1818	0.4807
D2	0.3076	1	0.1428	0.2222	0.5454	0.3076	0.2857	0	0.3333	0.2	0.3344
D3	0.5333	0.1428	1	0.3636	0.4615	0.4	0.5	0.4285	0.4285	0.1666	0.4425
D4	0.4	0.2222	0.3636	1	0.25	0.2	0.1818	0.2222	0.2222	0	0.3062
D5	0.5	0.5454	0.4615	0.25	1	0.6666	0.6153	0.3636	0.7272	0.2222	0.5352
D6	0.4285	0.3076	0.4	0.2	0.6666	1	0.6666	0.3076	0.7692	0.1818	0.4928
D7	0.5333	0.2857	0.5	0.1818	0.6153	0.6666	1	0.2857	0.7142	0.1666	0.4949
D8	0.4615	0	0.4285	0.2222	0.3636	0.3076	0.2857	1	0.3333	0.2	0.3602
D9	0.4615	0.3333	0.4285	0.2222	0.7272	0.7692	0.7142	0.3333	1	0.2	0.5189
D10	0.1818	0.2	0.1666	0	0.2222	0.1818	0.1666	0.2	0.2	1	0.2519

4. Overlap Similarity:

Overlap Similarity function was used as fourth similarity function in the experiment. In the experiment the first document terms were compared with terms of first document using overlap similarity function to obtain the value of overlap similarity as Olp1 and the terms of first document were compared with the terms of second document to obtain the value of Overlap similarity as Olp2 and so on. The value of overlap was obtained as Olp1, Olp2, Olp3, Olp4, Olp5, Olp6, Olp7, Olp8, Olp9 and Olp10. Then the average of all the coefficients was obtained. The above process is repeated for all other documents and the average of all the coefficients is taken as shown in table 4. The Overlap coefficient is a similarity function related to the Jaccard similarity function that computes the overlap between two sets. If set *X* is a subset of *Y* or the converse then the overlap coefficient is equal to one [18]. It falls in the category of term-based similarity measure. The formula for the overlap similarity function is given below:

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

TABLE4. OVERLAP SIMILARITY FOR QUERY "TERRORIST ATTACK MUMBAI"

Doc	Similarity with Overlap Similarity Function										Avg.
	Olp1	Olp2	Olp3	Olp4	Olp5	Olp6	Olp7	Olp8	Olp9	Olp10	
D1	1	0.3333	0.5714	0.6666	0.6	0.4285	0.5714	0.5	0.5	0.25	0.5421
D2	0.3333	1	0.1666	0.3333	0.6	0.3333	0.3333	0	0.3333	0.25	0.3683
D3	0.5714	0.1666	1	0.6666	0.6	0.4285	0.5	0.5	0.5	0.25	0.5183
D4	0.6666	0.3333	0.6666	1	0.3333	0.3333	0.3333	0.3333	0.3333	0	0.4333
D5	0.6	0.6	0.6	0.3333	1	0.8	0.8	0.4	0.8	0.25	0.6183
D6	0.4285	0.3333	0.4285	0.3333	0.8	1	0.7142	0.3333	0.8333	0.25	0.5454
D7	0.5714	0.3333	0.5	0.3333	0.8	0.7142	1	0.3333	0.8333	0.25	0.5669
D8	0.5	0	0.5	0.3333	0.4	0.3333	0.3333	1	0.3333	0.25	0.3983
D9	0.5	0.3333	0.5	0.3333	0.8	0.8333	0.8333	0.3333	1	0.25	0.5716
D10	0.25	0.25	0.25	0	0.25	0.25	0.25	0.25	0.25	1	0.3

D. Average Similarity Values for the different queries

In the experiment the process explained above was repeated for the ten different queries and similarity score was measured by using the above said formulae of Jaccard Similarity, Cosine Similarity, Dice Similarity and Overlap similarity functions .Average similarity value was measured for the ten obtained values of similarity. The results obtained from the experiment are summarized in the table 5 and are shown in fig.1.

TABLE 5 AVERAGE SIMILARITY VALUES FOR DIFFERENT QUERIES USING JACCARD, COSINE, DICE AND OVERLAP SIMILARITY FUNCTIONS

Query No.	Query Entered in Search Engine	Jaccard Similarity	Cosine Similarity	Dice Similarity	Overlap Similarity
Q1	Terrorist Attack Mumbai	0.3111	0.4280	0.4218	0.4863
Q2	Cloud Burst India	0.2277	0.3112	0.3085	0.3427
Q3	Moist Attack India	0.2443	0.3345	0.3262	0.3960
Q4	Corruption Cricket India	0.2906	0.4093	0.4047	0.4592
Q5	Pollution River Ganga	0.4493	0.5969	0.5914	0.6645
Q6	Power Generation India	0.2800	0.3823	0.3784	0.4269
Q7	Sand Mining India	0.3898	0.5210	0.5176	0.5675
Q8	Mid Day Meal India	0.3111	0.4278	0.4198	0.4949
Q9	Sikh Riots India	0.3536	0.4784	0.4763	0.5141
Q10	Moist Attack Train	0.3760	0.5116	0.5070	0.5627

From the table 5and fig.1, it is clear that the average Overlap similarity outperforms the average Cosine similarity which in turn outperforms the average Dice similarity and the average Jaccard similarity for the ten different queries i.e. Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9 and Q10. It is also concluded that the results obtained from the Cosine similarity function and the Dice similarity functions are quite comparable. The average Overlap similarity is highest for each query and average Jaccard similarity is lowest for each query. The average similarity measured using the four similarity functions is lowest for query, Q2 and the average similarity measured using the four similarity functions is highest for the query, Q5.

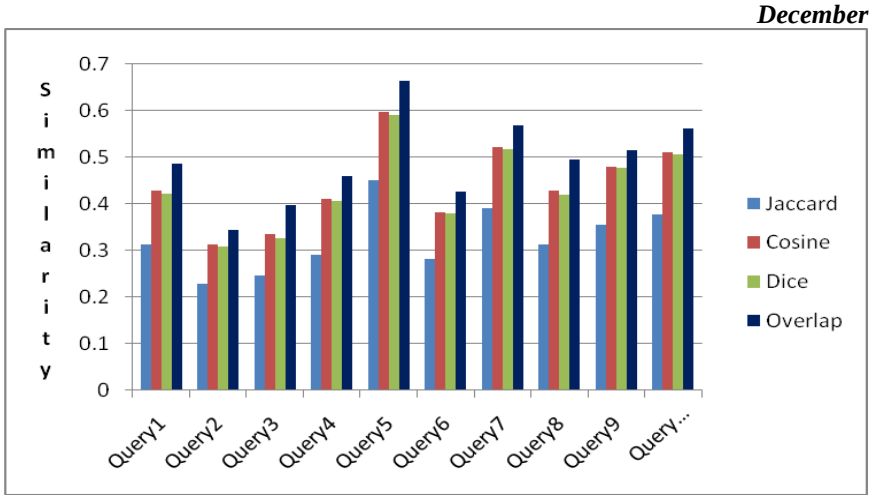

Fig.1 Average similarity of different queries with the different similarity functions

VI. ANALYSIS AND CONCLUSIONS

In this paper the textual similarity of the text entered for the different queries is measured using the four similarity functions i.e. Jaccard, Cosine, Dice and Overlap. Fig.1 explains that the similarity of Overlap coefficient outperforms the Cosine similarity measure and Cosine similarity is more than the Dice and Jaccard similarity for the different adhoc queries. Different similarity functions have been proposed in the field of information retrieval in wide area networks and it is concluded that the textual similarity between the document and query can be improved either by the proper design of similarity measures or by modifying the query or enhancing the query because the retrieval process is iterative search process in which the user interacts iteratively with the search system for the information need.

REFERENCES

[1] (2013) [Online]. Available: http:// www.internetlivestats.com/internet-users/

[2] James Allan, Bruce Croft, "Challenges in Information Retrieval and Language Modeling," Univ. of Massachusetts, Amherst, Center for Intelligent Information Retrieval, Technical Report, 2002.

[3] M. Kobayashi and K. Takeda, "Information Retrieval on Web," *ACM Computing Surveys*, Vol. 32, No.2, pp.144-168, 2000.

[4] Sergey Brin and Lawrence Page, "The Anatomy of Large-Scale Hyper textual Web Search Engine," *Proc. of 7th International World Wide Web Conference/Computer Networks Vol.1-7*, pp.107-117, 1998

[5] Dan Farber (2013) The CNET website [Online]. Available:http://www.cnet.com/google-search-scratches-its-brain-500-million-times-a-day/

[6] Nicholas J. Belkin, W. Bruce Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" *Communications of the ACM*, Vol.35, No.12, p29 (10), 1992.

[7] Nicholas J. Belkin , W. Bruce Croft, "Retrieval Techniques," *Annual Review of Information Science & Technology*, M.E Williams, Ed. Chapter. 4, pp.109-145 Elsevier, 1987.

[8] G. Salton, M. H. McGill, *Introduction to Modern Information Retrieval*, McGraw Hill, 1983.

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

[10] Michael Gordon, "Applying Probabilistic and Genetic Algorithms for Document Retrieval," Computer Practices, 1208-1218, 1988.

[11] William P. Jones, George W.furnas, "Picture of Relevance: A Geometric Analysis of Similarity Measures," *Journal of the American Society for Information Science*, Vol.38, No.6, pp.420-442, 1987.

[12] Sung-Hyuk Cha, "Comprehensive Survey on the Distance/Similarity Measures between Probability Density Functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, Vol. 1, Issue 4, pp. 300-307, 2007.

[13] Siti Salwa Salleh, Noor Aznimah Abdul Aziz, Daud Mohamad and Megawati Omar, "Combining Mahalanobis and Jaccard Distance to Overcome Similarity Measurement Constriction on Geometrical Shapes," *International Journal of Computer Science Issues*, Vol. 9, Issue 4, pp. 124-132, 2012

[14] E. Kosman and K. J. Leonard, "Similarity Coefficients for Molecular Markers in Studies of Genetic Relationships between the Individuals for Haploid, Diploid and Polyploidy Species," *Molecular Ecology*, Vol. 14, Issue 2, pp. 415-424, 2005.

[15] Jair Moura Duarte, Joao Bosco dos Santos and Leonardo Cunha Melo, "Comparison of Similarity Coefficients Based on RAPD Markers in the Common Bean," *Genetics and Molecular Biology*, Vol. 22, Issue 3, pp. 427-432, 1999.

[16] P. Wallet, J. M. Barnard and G.M. Downs, "Chemical Similarity Searching," *Journal of Chemical and Information and Computer Sciences*, Vol. 38, No. 6, pp. 983-996, 1998.

[17] McGill, Koll and Noreault, "An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems," Project report, Syracuse University, 1979.

[18] Wael H. Gomaa, Aly A. Fahmy, "A Survey of Text Similarity Approaches," *International Journal of Computer Applications*, Vol. 68, No. 13, pp. 13-18, 2013.

[19] Suphakit Niwattanakul, Jatsada Singhthongchai, Ekkachai Naenudorn and Supachanun Wanapu, "Using of Jaccard Coefficient for Keywords Similarity," *Proc. of International MultiConference of Engineers and Computer Scientists*, Vol.1, IMECS 2013, 2013, Hong Kong.

[20] Wa'el Musa Hadi, Fadi Thabtah, Hussein Abdel-jaber, "A Comparative Study Using Vector Space Model with K-Nearest Neighbor on Text Categorization Data," *Proc. of the World Congress on Engineering, WCE2007*,2007,Vol.1, London, U.K.

[21] Xiaojun Wan, "A Novel Document Similarity Measure based on Earth Movers Distance," *Information Sciences*. 177, pp. 3718-3730, 2007.

[22] http://textalyser.net.

[23] http://in.mathworks.com/products/matlab./

