ISSN:0975-9646

A Study of Similarity Functions Used in Textual Information Retrieval in Wide Area Networks

Jaswinder Singh [#1], Parvinder Singh [*2], Yogesh Chaba [#3]

[#] *Department of Computer Science & Engineering, Guru Jambheshwar University of Science & Technology Hisar, Haryana, India*

[*] *Department of Computer Science & Engineering, Deenbandhu Chhotu Ram University of Science & Technology Murthal, Sonepat, Haryana, India*

***Abstract*— World Wide Web is a rich source of information. It continues to expand in size and complexity with the increasing use of the internet and social media but how to retrieve relevant documents on the Web is becoming a challenge. In this paper there is discussion about the goals, challenges and importance of similarity functions in information retrieval in wide area networks. This paper discusses the different similarity functions that are used by various authors as information retrieval techniques to measure the similarity of document with the query in the field of information retrieval in wide area networks.**

***Keywords*— Similarity Function, Textual Information Retrieval, Wide Area Networks**

I. INTRODUCTION

The continuous growth of web and the expectation of user on search engine to anticipate his or her needs have led to the development of the field of information retrieval in wide area networks. The tool used to extract relevant information from web world is called search engine. A survey claim that 85% of internet users use search engines or some kind of search tool to find specific information of interest [1]. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. Search engines answer tens of millions of queries every day [2]. The objective of search engine is to provide high quality results to the user that is relevant to the user query. Search engines use automated software programs known as spiders to survey the web and build their database. Web documents are retrieved by these programs are analysed. Data collected from each web page are then added to the search engine index. 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 URLs are then returned to the user as hits, ranked in order with the best results at the top. The aim of this paper is to study the goals, challenges and importance of similarity functions in the field of information retrieval in the
wide area networks, particularly the similarity functions. The remainder of paper is organized as follows. The first section of paper describes the brief working of search engine and second section describes the information retrieval process in wide area networks. This section describes the goals, challenges of information retrieval and the problems that are faced by information retrieval system in wide area networks that is whether because of the nature of web or because of the activity of user or the searching process. This section also describes the information retrieval system and the classical models of information retrieval in wide area networks. Third section describes the various similarity functions which are the functions that are used to find out the textual similarity between the user query and documents. The related work on the similarity functions is reviewed and concludes that with the proper combination of the similarity functions the search results can be further improved.

II. INFORMATION RETRIEVAL IN WIDE AREA NETWORKS

Information retrieval has become an important subject of much research in recent years, because the amount of information available in digital formats has grown exponentially and the need for retrieving relevant information has assumed a crucial importance. The most common text retrieval task is to retrieve the documents in response to the user query. "Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information" [3]. Information Retrieval deals with representation, storage, organization of, and access to information items such as documents, web pages, etc [4]. Information Retrieval system is different from the DBMS in the sense that the retrieval is probabilistic where as the retrieval is deterministic in DBMS [5]. Modern information retrieval can be accessed through the services of different search engines e.g. Google, Bing and AltaVista etc.

A. Goals of Information Retrieval

The main goal of an Information Retrieval in wide area networks is to search for the documents that are relevant to the user's query. Keyword search is the simplest form of the most popular query method for the search engine in information systems. Searched results of inputted keyword in some cases might not display the required documents. This can be the result of lacking of search method or knowledge of how to use the specific keyword. Fig.1 explains the information retrieval process which is mostly followed by the user during searching the information in wide area networks. User formulates a query about the information need and then the user chooses the search tool or search system and sends to the information retrieval system. Information retrieval system searches for the matches in the document database and retrieves the results. The user evaluates the results based on the relevance [4]. Relevance is subjective in nature as it depends on the

judgment of users. Goal of search component is to predict which documents are relevant to the user need and rank the documents in the order of predicted likelihood of relevance of user. The documents with more similarity with the user's query will have high relevancy and be at higher position in the retrieved documents list. The relevancy can be measured by the similarity between the documents and query by using the similarity functions [4]

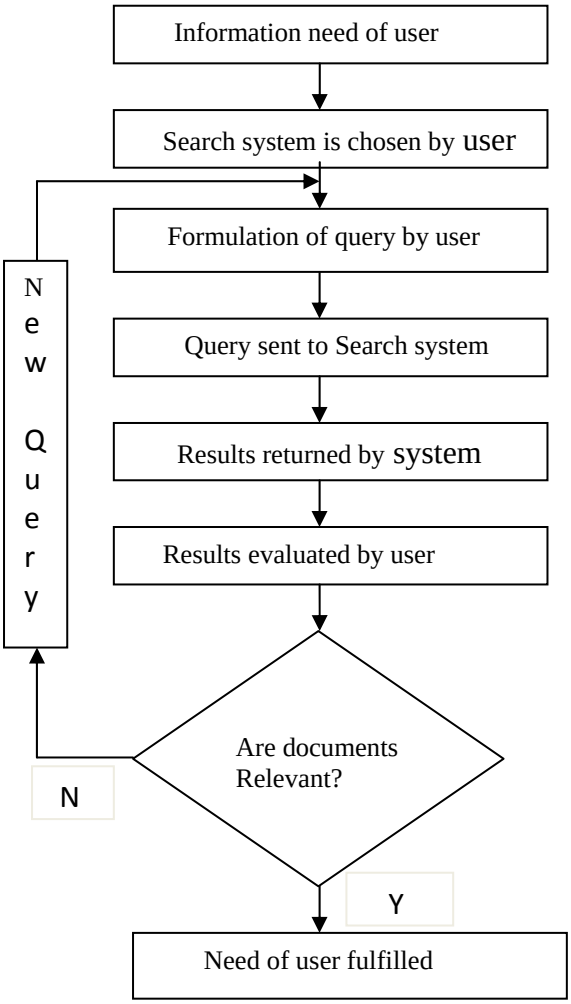

Fig.1 Information Retrieval Process in Wide Area Networks

B. Information Retrieval Challenges

As the web keeps growing in size, the problem of searching the web is becoming more complex. There are different information retrieval challenges in the wide area networks. These challenges can be categorized as challenges of the web, problems with data itself, types of user searching the information and the searching process.

Expansion of Web:

With the expansion of the Web, there is an increase in volume of data and information published in various Web pages.

Language Problem:

Most of the information on the web is in English. So the person who does not know English, the vast information is not available. This problem is more appropriate for the multi-lingual and multi-cultural country like India.

Variety of data:

There is variety of data source. Information of almost all types exists on the Web. e.g. texts, image , videos , songs , photographs & multimedia data etc.

Data is distributed:

Due to the intrinsic nature of Web, data is distrusted over many computers and platforms.

Information is linked:

Most of the information on Web is linked.

Redundancy:

Much of information on the web is redundant. There may be distribution of one or more copies of the same information on the web. Several studies indicate that nearly 30% of contents of web is duplicated [6], [7].

Noise in web:

A Web page contains different kinds of information as with the display of main contents the page may contain advertisement etc.

Dynamics:

There is freedom for anyone to publish information on the web at anytime and any where implies that information on the web is constantly changing. It is estimated that 40% of the web changes every month [4].

Data Quality:

Data can be false or invalid as it poorly written or have many errors as web is considered as new source of publishing i.e there are so many journals and in most of cases, no editorial process is followed.

Search behaviour of users:

The users using the web have different search behaviour. They have different expectations and goals such as Informative, Transactional and Navigational [8]. The search query which is entered with the intent of finding the particular website or web page is called the Navigational search query. The query which covers the broad topic for which there may be thousands of results is known as the informational search query. The query that indicates intent to complete the transaction such as making a purchase, downloading a file is known transactional search query. Based on their search strategies the internet users are broadly classified into three categories i.e. a casual user searching the web for something that is loosely defined, a researcher looking for serious research content over the web, a professional looking for business intelligence by searching the web.

In sufficient query:

The user queries are limited to few keywords and the user often does not know the best query to retrieve the information need.

Judgment of retrieved documents:

For the given query the same document may be judged as relevant by the user and non relevant by another user.

C. Information Retrieval Models in Wide Area Networks

Information retrieval model is a pattern that defines several aspects of a retrieval procedure i.e. how the documents and user's query are represented and how a system retrieves the relevant documents according to the user's query and how the retrieved documents are ranked. Information retrieval techniques can be categorised as the exact match and the partial match retrieval techniques. The exact match retrieval

techniques will retrieve the documents which exactly match the query and the partial match retrieval techniques will retrieve the documents that also include the documents that exact match query. Several studies have been proposed in the literature that classifies the IR models. These models are categorised as the classical models and the non classical models [9]. The non classical models of information retrieval models are based on the principles other than the similarity, probability and Boolean operations etc. These include the information logic model and interaction model. Other alternative models of information retrieval include the cluster model and the fuzzy model and latent semantic indexing model. Gerado Canfora et.al [10] proposed taxonomy of information retrieval models. The vertical taxonomy classifies information retrieval models based on a two components view, namely representation and reasoning. The horizontal taxonomy classifies information retrieval objects with respect to the application areas. Three types of classical models of information retrieval in wide area networks are defined [3].These models have formed the basis of information retrieval research in wide area networks.

Boolean Model: Boolean model is based upon the Boolean logic and classical sets theory. Retrieval is based on whether or not the documents contain the query terms. The Boolean model is very rigid: AND means "all", OR means "any". In Boolean model the similarity function is Boolean and retrieved documents are not ranked. Boolean operator usage has much influence. In general the Boolean model is considered as the weakest model its main weakness is inability to recognize the partial match. The documents that exactly match the query are retrieved. Boolean model is still used for small scale search like searching the file on hard drive or e-mails etc.

Probabilistic model: probabilistic model ranks the documents based upon the probability of their relevance to the given query [4].The probabilistic model in its pure form have been implemented with small scale search tasks like library catalogue works [8].

Vector Space Model: The vector space model is the most well studied retrieval models. The important contributions to its development were made by Lunh [11], Salton [3], Salton and McGill [12] and Van Rijsbergen [13]. It generates weighted term vectors for each document in the collection, and for the user query. The retrieval is based on the similarity between the query vector and document vectors. Based upon the similarity the output documents are ranked accordingly. The term is weighted with importance. Partial match is there. The similarity is based on the occurrence frequencies of the keywords in the query and in the document. The main disadvantage is that it assumes that the terms are independent. The vector space model is dominant thought among the researchers, practitioners and web community, where the popularity of vector space model runs high .This model outperforms probabilistic models in large scale information retrieval tasks [4]. So an information retrieval model in wide area networks consists of representation for documents, representation for queries, a modelling framework for the documents and queries and relationship between them, a ranking or similarity function which orders the documents with respect to query.

D. Information Retrieval System

The basic component of any information system is the representation of the information itself. In the text information retrieval, representation means the representation of documents and queries. Representation of queries means the representation of user need. An information retrieval system is defined as a system which interprets the contents of information items and generates ranking which reflect relevance and retrieves the information more efficiently. The keywords are used by the most of Information retrieval system to retrieve 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: Documentary Database, Query Subsystem, Matching mechanism [4], [22].This document database stores document along with the information content of their representation. It is associated with the indexer module which automatically generates are presentation for each document by extracting the document contents. Query subsystem is a system which formulate user request into query. Matching function compares the similarity between the query and document in the database. Based on this, documents are retrieved. The success of any information retrieval system depends on the ability to access the relevance of objects in its database to a given user's request [14]. The effectiveness of retrieval system can be explained by two parameters used for many years i.e. recall and precision. Recall is the ratio of the number of relevant documents retrieved to the number of relevant documents in the collection. Precision is the ratio of number of relevant documents retrieved to the total number of documents retrieved [7].

III. SIMILARITY FUNCTIONS USED IN INFORMATION RETRIEVAL IN WIDE AREA NETWORKS

From the literature it was found that there are many similarity functions which are used in the various fields such as information retrieval [15], image retrieval [16], molecular ecology [17], genetics and molecular biology [18] and chemistry [19]. In the information retrieval, similarity functions are used as the information retrieval techniques and similarity functions are the functions which are used to measure the 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. It is therefore difficult to decide which measure should be used for a particular application. The techniques which are used to measure the similarity between the documents and query are the textual

information retrieval techniques which make use of similarity functions .The similarity functions which are used by the various authors in the field of information retrieval in wide area networks are explained below.

Inner product :

One of the most commonly used similarity function is to take the inner product between the query and document vector. Similarity between vectors for the document ***dj*** and query ***qk*** can be computed as the vector inner product.

$$\text{Sim}(\boldsymbol{dj}, \boldsymbol{qk}) = \boldsymbol{dj} \bullet \boldsymbol{qk} = \sum_{i=1}^{m} w_{ij} w_{ik}$$

Where w_{ij} is the weight of term i in document j , w_{ik} is the weight of term i in the query k and m is the no of terms used to represent documents in the collection.

•For binary vectors, the inner product is the number of matched query terms in the document (size of intersection).

•For weighted term vectors, it is the sum of the products of the weights of the matched terms.

Jaccard Similarity Function:

The Jaccard similarity function is defined as the size of the intersection divided by the size of the union of the document and query vectors as expressed below

$$= \frac{\sum_{i=1}^{m} w_{ij} w_{ik}}{\sum_{i=1}^{m} w_{ij}^2 + \sum_{i=1}^{m} w_{ik}^2 - \sum_{i=1}^{m} w_{ij} w_{ik}}$$

Dice Similarity function:

Dice coefficient is defined as twice the number of common terms in the compared strings divided by the total number of terms in both strings.

$$\text{Sim}(\text{dj}, \text{qk}) = \frac{2 \sum_{i=1}^{m} w_{ij} w_{ik}}{\sum_{i=1}^{m} w_{ij}^2 + \sum_{i=1}^{m} w_{ik}^2}$$

Cosine Similarity function:

Cosine formulation measures cosine of the angle between the query and document vector .The numerator of the cosine measure is itself nothing more than the inner product measure. The critical difference, then, is that the inner product is divided by the product of Euclidean lengths of the document and query vectors.

$$\text{Sim}(\text{dj}, \text{qk}) = \frac{\sum_{i=1}^{m} w_{ij} w_{ik}}{\sqrt{\sum_{i=1}^{m} w_{ij}^2} \sqrt{\sum_{i=1}^{m} w_{ik}^2}}$$

Overlap Similarity Function:

The overlap similarity function is obtained as shown below

$$\text{Sim}(\text{dj}, \text{qk}) = \frac{\sum_{i=1}^{m} w_{ij} w_{ik}}{min(\sum_{i=1}^{m} w_{ij}, \sum_{i=1}^{m} w_{ik})}$$

IV. Importance of Similarity Functions Used in Information Retrieval in Wide Area Networks

The vast increase in the amount of online text and the need of the different types of information have led to the interest in the different areas of information and retrieval like multimedia retrieval, chemical and biological informatics, topic detection and summarization etc. [20]. Despite this, Textual similarity is the basis of all the above said fields. Textual similarity functions plays the important role in the text related research and the tasks related to its applications in the field of information retrieval, topic detection, text classification. The textual similarity makes use of similarity functions. The textual similarity function is partitioned into String-based, Corpus-based and knowledge-based. String-based is further characterized as the character-based approach and the term based approach [21].The term-based approach makes use of Jaccard, Cosine, Dice and Overlap similarity functions. If binary weights are used, then weight of term can be 1 if term occurs in the document and 0 if the term does not occurs in the document then all the stated formulae of section III of the paper for the similarity function in the binary term vectors are shown in the Table1.

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

Y is defined, a 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.

TABLE I Similarity Functions with Binary Weights

Sr. No.	Similarity Function	Similarity with Binary term vector
1	Inner Product	$\|X \cap Y\|$
2	Jaccard	$\frac{\|X \cap Y\|}{\|X\| + \|Y\| - \|X \cap Y\|}$
3	Dice	$2\frac{\|X \cap Y\|}{\|X\| + \|Y\|}$
4	Cosine	$\frac{\|X \cap Y\|}{\|X\|^{1/2} . \|Y\|^{1/2}}$
5	Overlap	$\frac{\|X \cap Y\|}{min(\|X\|, \|Y\|)}$

There are number of similarity functions found in literature and above five formulae of similarity functions which were defined in Table1 were used frequently as information retrieval techniques in the field of information retrieval in wide area networks.

V. CONCLUSIONS

In this study different similarity functions that are used as information retrieval techniques in wide area networks to measure the similarity between the documents and query were discussed i.e. Inner product, Jaccard, Dice, Cosine and Overlap similarity functions. All of these similarity functions fall in the category of term based similarity functions which is sub category of the string based similarity functions. It is also concluded that the vector space model is most dominant thought among the researchers and web community where the retrieval process is based on the similarity between the query vector and document vectors. The weights can be raw term weights or the binary weights where the raw weights are the frequency of occurrence of term in each document and binary weights are the presence or absence of terms. Based upon the numeric similarity between the query and document, the documents can be ranked. Different similarity functions have been proposed in the field of information retrieval in wide area networks and it is concluded that with the proper combination of similarity functions the similarity between the document and query, the textual similarity can be enhanced.

REFERENCES

[1] M. Kobayashi and K. Takeda, "Information Retrieval on the Web," *ACM Computing Surveys*, Vol.32, No.2, 2000.

[2] Sergey Brin and Lawrence Page, "The Anatomy of Large Scale Hyper textual Web Search Engine," in *Proc. 7th International World Wide Web Conference, Computer Networks and ISDN Systems*, Vol. 30, Issue 1-7, pp.107-117, 1998

[3] G. Salton, Automatic Information Organization and Retrieval, McGraw-Hill, New York, 1968.

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

[5] William B. Frakes, R. Baeza-Yates, Information Retrieval Data Structure and Algorithms, Pearson, 2008.

[6] M.P.S. Bhatia, Akshi Kumar Khalid,, "Information Retrieval and Machine Learning: Supporting Technologies for Web Mining Research and Practice," *Webology*, Vol.5, No.2, 2008.

[7] M.P.S. Bhatia, Akshi Kumar Khalid,, "A Primer on the Web Information Retrieval Paradigm," *Journal of Theoretical and Applied Information Technology Vol.4, No.2*, pp.657-662, 2008.

[8] C. D. Manning, P. Raghavan and H.Schutze , An Introduction to Information Retrieval, Cambridge University Press, 2009.

[9] Tanveer Siddiqui, U.S.Tiwari, Natural Language Processing and Information Retrieval, Oxford University Press, India, 2008.

[10] Gerardo Canfora and Luigi Cerulo, "A Taxonomy of Information Retrieval Models and Tools," *Journal of Computing and Information Technology*, Vol.12, No.3, pp.175-194, 2004.

[11] H.P Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal of Research and Development*, Vol. 1, Issue 4, pp.309-317, 1957.

[12] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[13] C.J. Van Rijsbergen, "A Theoretical Basis for the Use of Co-occurrence data in Information Retrieval," *Journal of Documentation*, Vol. 33, Issue 2, pp.106-119, 1977.

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

[15] 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.

[16] 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

[17] 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.

[18] 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.

[19] 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.

[20] James Allan, Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan , Bruce Croft , "Challenges in Information Retrieval and Language Modeling," Univ. of Massachusetts, Amherst, Center for Intelligent Information Retrieval Technical Report, 2002.

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

[22] Praveen Pathak, Michael Gordon, Weiguo Fan, " Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaption" in *Proc. International Conference on System Sciences*, Hawaii, USA, 2000.