International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 7, July - 2014

Relevancy Measurement of Retrieved Webpages Using Ruzicka Similarity Measure

Manjeet*, Jaswinder Singh**
*Master of Technology (Dept. of Computer Science and Engineering) GJUS&T, Hisar, Haryana, India
**Assistant professor (Dept. of Computer Science and Engineering) GJUS&T, Hisar, Haryana, India

***Abstract*— The information in the web is increasing day by day causing the information retrieval task more and more difficult. Search engines are used to retrieve the information these days. The goal is to retrieve most relevant documents with less number of irrelevant documents with respect to user's query in information retrieval system. Similarity Measures is a function that is used to measure the degree of similarity between query and documents. It measures how much the query and document is similar with each other. In this paper the relevancy of the retrieved web pages is calculated after submitting the query into the search engine.**

Keywords*— Vector Space Model, Information Retrieval, Similarity Measure.

1. INTRODUCTION

Information retrieval (IR) is the process of extracting material in an unstructured way that satisfies an information need from within large collections usually stored on computers [1]. An information retrieval process begins when a query is entered into the system. There are three basic processes an information retrieval system needs to support: the representation of the contents of the documents, the representation of the information needed by the user, and the comparison of these two representation methods. The processes are visualized in Fig.1. In the figure, square boxes represent data and round boxes represent processes. The process of representing the information need is referred to as the formulation of the query. The resulting representation is the query. Representing the documents is usually called the indexing process. The comparison of the query against the document representations is called the matching process [2].

A. Components of Information Retrieval System

There are three basic components of the Information Retrieval System. They are Documentary Database, Query Subsystem and matching function.

I. Query subsystem

Query subsystem is a system which allows the users to formulate their queries and present the relevant documents retrieved by the system.

II. Matching function

Matching functions compare both query and documents in the database and return a value which measures the similarity between query and the documents. With the help of this, relevant documents from the database are retrieved.

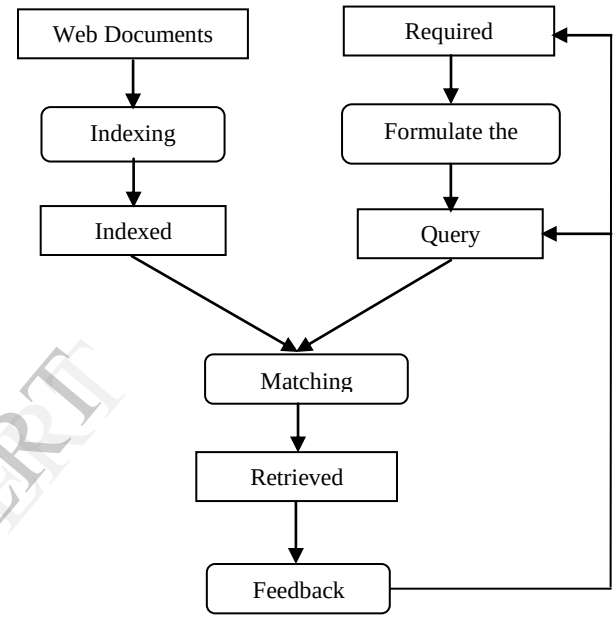


Fig. 1

III. Document database

It is the storage where the documents are stored. It also represents the information contents of the documents. Matching Function compares all the documents of document database with the query and extracts relevant documents.

B. Information Retrieval Models

I. Boolean Model

In this model the indexer module performs a binary indexing. In other words a term in a document representation is either significant or not. User's queries in this model are expressed using a query language that is based on these terms and allow combination of simple user requirements with the logical operators. The result obtained by the processing the query is a set of documents that completely match with it. Only two possibilities are considered for each document i.e. to be or not to be relevant for the user's needs.

II. Vector Space Model

In Vector Space Model, a document is viewed as a vector in n-dimensional document space. The query is also treated in the same way and constructed from the terms and weights provided in the user's request. Document retrieval is based on the measurement of the similarity between the query and the documents. This means that documents with higher similarity to the user's query will be retrieved in a higher position in the list of retrieved documents. In this way, the retrieved documents are orderly presented to the user with respect to their relevancy to the query.

III. Probabilistic Model

This model uses the probability theory to build the search function and the operating mode of it. The information used to compose the search function is obtained from the distribution of the index terms throughout the collection of documents or from its subset. This information is used to set the values of some parameters of the search functions, which are composed of a set of weights associated to the index terms [3].

C. Similarity Measures

Similarity Measures is a function that is used to measure the degree of similarity between query and documents. It measures how much the query and document is similar with each other. It returns a value which decides the degree of similarity. In order to find the similarity query and document are converted into vector form. Types of similarity measures are discussed as follows:

I. Distance-Based Similarity Measures

It is one of the oldest and most influential assumption that the similarity between the query and the documents is inversely proportional to the psychological distance.

II. Feature-Based Similarity Measures

In 1977 Tversky proposed that the similarity is the result of the process of feature matching. This process differentially weights the same and different features.

III. *Probabilistic Similarity Measures*

These models make two assumptions. Some probabilistic models assume that similarity is inversely related to psychological distance [4].

Some most frequently used similarity Measures are:

I. Cosine

$$S_{Cos} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sqrt{\sum_{i=1}^{d} P_i^2}\sqrt{\sum_{i=1}^{d} Q_i^2}}$$

II. Jaccard

$$S_{Jac} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}$$

III. Dice

$$S_{Dice} = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}$$

IV. Ruzicka

$$S_{Ruz} = \frac{\sum_{i=1}^{d} \min(P_i, Q_i)}{\sum_{i=1}^{d} \max(P_i, Q_i)}$$

D. Page Relevancy

Relevancy Measurement is the process of measuring that the retrieved web pages are how much percent related to the information requested by the user. The documents with high similarity with the user's query will return high relevancy value and must be at higher position in the retrieved documents list. The main aim is to remove the irrelevant web pages by measuring their relevancy with the user's query.

2. PREVIOUS WORK

M. C. McCabe et al. [5] presented that fusion of various retrieval strategies is a means of improving retrieval effectiveness. The effect of fusion on various query representations has shown a twelve percent improvement in average precision. *Bangorn Klabbankoh et al.* [6] presented that under vector space model, information retrieval is based on the similarity measurement between query and documents. *Shyam Boriah et al.* [7] studied the performance of a variety of similarity measures in the context of a specific data mining task: outlier detection. Results on a variety of data sets showed that while no measure dominates others for problems of all types, some measures have consistently high performance. *Shini Renjith* et al.[8] proved that cosine similarity measure is Jaccard Coefficient and Inner Product. It has been proved that the cosine coefficient is the best among the above discussed coefficients. Precision and recall were taken as the measures for evaluating the efficiency of the coefficients. *J. Usharani et al.* [9] presented method to find the similarity of web documents based on cosine similarity. *Pragati Bhatnagar et al.* [10] described the concept of Information Retrieval System (IRS), Evolutionary algorithm and the appropriateness of evolutionary algorithm to retrieve the relevant information. In the later sections the proposed model for adaptive Information Retrieval System and algorithm for implementing proposed system is presented. *Andrei Z. Broder, et al.* [11] presented an efficient query evaluation method based on a two level approach: at the initial level, it iterates in parallel over query term postings and identifies candidate documents using an approximate evaluation taking into account only partial information on term occurrences and no query independent factors and at the second level, promising candidates are evaluated and their exact scores are computed. The investigation demonstrates that using the document at- a-time approach and a two level query evaluation method using the wand operator for the first stage, pruning can yield substantial gains in efficiency.

A*shish Kishor Bindal et al.* [12] described a stochastic based approach for optimizing query vector without user involvement. *Abdelmgeid A. Aly* [13] presented an adaptive method to modify user's queries, based on relevance judgments. The main aim was to provide most relevant web pages and to minimize the irrelevant information. *Sonali Sonksusare et al.* [14] attempted to make a survey on the techniques used for the retrieval of information in order to remove the errors from the old retrieval techniques. *A. Haritha et al.* [15] traversed different hyperlinks to provide more relevant pages and transformed them to documents and used the numerical measures like Euclidian distance and Cosine similarity to measure the orientation of the websites to each other. *Manoj Chahal et al.*[16] used the Horng and Yeh coefficient to increase the efficiency of Information Retrieval System. *Seung-Seok Choi et al.* [17] described that the binary feature vector is one of the most common representations of patterns. A few comprehensive surveys on binary measures have been conducted. Hence 76 binary similarity and distance measures used over the last century are collected and revealed their correlations through the hierarchical clustering technique. *Sung-Hyuk Cha* [18] showed that distance or similarity measures are essential to solve many pattern recognition problems such as classification, retrieval and clustering problems. Various distance or similarity measures, that can be applied to compare two probability density functions are reviewed and categorized in both syntactic and semantic relationships. *Md. Abu Kausar et al.* [19] dealt with the basics of the information retrieval. The research areas in web search is discussed in this paper. It also deals with the different proposals in web search which are promising research areas. A review of the research works done in information retrieval domain has been discussed. *Hazra Imran* et al. [20] suggested that the central problem of information retrieval is to measure the relevancy of the documents with the information needed by the user. The selected similarity measure plays a crucial role in improving search effectiveness of a retrieval system.

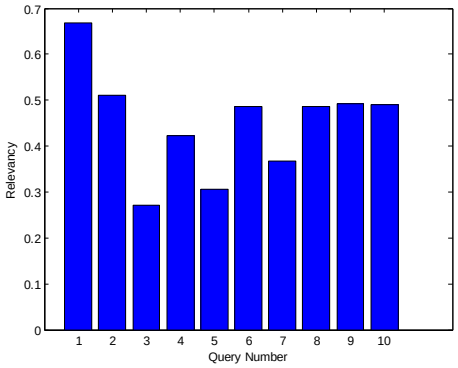
From the review of literature, factors that affects the performance of query expansion methods have been identified. Since there exists multiple retrieval models with benefits and limitations, the impact of retrieval models on the performance of query expansion is important.

3. EXPERERIMENT

Different similarity measures have been suggested to match the query and documents. We used Ruzicka Similarity Coefficient to measure the relevance of web page. First of all the query was submitted to the search engine and relevant documents were retrieved. Then the chromosomes were generated according to the keywords present in the documents. The length of each chromosome selected was 25 digits. After the chromosome generation the relevancy of each individual document was calculated according to the RUZICKA Similarity Measure.

After this the average relevancy of the documents with the query is calculated which is shown by the relevancy table and accordingly bar graph is drawn.

Sr. No.	Query No.	Average Relevancy
1.	Query 1	0.6690
2.	Query 2	0.5112
3.	Query 3	0.2702
4.	Query 4	0.4225
5.	Query 5	0.3058
6.	Query 6	0.4857
7.	Query 7	0.3682
8.	Query 8	0.4858
9.	Query 9	0.4932
10.	Query 10	0.4906



4. CONCLUSION AND FUTURE WORK

Query expansion is a process that aims to reformulate a query to improve the results of information retrieval. It involves adding new words and phrases to the existing search terms to generate an expanded query. We measured the relevancy of the retrieved web pages after submitting the query into the search engine and a bar graph is drawn which shows the average relevancy of the documents retrieved with the query. After calculating the relevancy of the documents Various GA operators (Crossover, Mutation) can be applied to find new chromosomes. These will further undergo the same process using GA and the change in the relevancy can be obtained.

5. REFERENCES

1. C. D. Manning, P. Raghavan, and H. Schütze, Introduction to information retrieval, vol. 1. Cambridge university press Cambridge, 2008.
2. D. Hiemstra, "Information retrieval models," Information Retrieval: searching in the 21st Century, pp. 1–19, 2009.
3. Priya I. Borkar, Assistant Prof. Leena H. Patil, "A Model of Hybrid Genetic Algorithm-Particle Swarm Optimization (HGAPSO) Based Query Optimization for Web Information Retrieval" IJRET | Volume: 2 Issue: 1.JAN 2013.
4. http://www.scholarpedia.org/article/Similarity_measures.
5. M. C. McCabe, A. Chowdhury, D. Grossman, and O. Frieder, "System fusion for improving performance in information retrieval systems," in Information Technology: Coding and Computing, 2001. Proceedings International Conference on, 2001, pp. 639–643.

6. B. Klabbankoh and Q. Pinngern, “Applied genetic algorithms in information retrieval”, Faculty of Information Technology, King Mongkuts Institute of Technology Ladkrabang, 2000.
7. S. Boriah, V. Chandola, and V. Kumar, “Similarity measures for categorical data: A comparative evaluation”, red, vol. 30, no. 2, p. 3, 2008.
8. Shini Renjith Anjali C , "Fitness function in Genetic Algorithm based information filtering" International Journal of Computer Science and Mobile Computing, ICMIC13, December- 2013, pg. 80-86.
9. J. Usharani, K. Iyakutti, “A Genetic Algorithm based on Cosine Similarity for Relevant Document Retrieval” International Journal of Engineering Research & Technology, Vol. 2 Issue 2, February- 2013.
10. Pragati Bhatnagar, N. K. Pareek, “A Combined Matching Function based Evolutionary Approach for development of Adaptive Information Retrieval System”, International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 6, June 2012.
11. Andrei Z. Broder, David Carmel, Michael Herscovici, AyaSoffer, Jason Zien,“Efficient Query Evaluation using a Two-Level RetrievalProcess”CIKM’03, November 3–8, 2003.
12. S. Sanyal and A. K. Bindal, “Query Optimization in Context of Pseudo Relevant Documents” in 3rd Italian Information Retrieval (IIR) workshop, 2012.
13. Abdelmgeid A. Aly, “Applying Genetic Algorithm in Query Improvement Problem” International Journal Information Technologies and Knowledge Vol.1 / 2007.
14. Sonali Sonksusare, Mr. Jayesh Surana, “A Survey on Different Techniques for Data Classification and Information Extraction from the Websites” International Journal of Advanced Research in Computer Science and Software Engineering 3(11), November - 2013, pp. 383-386.
15. A. Haritha, P.V.S. Lakshmi, “Usage of Similarity Measures to Cluster Related Web Links” International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 11, November 2013.
16. Manoj Chahal, Jaswinder Singh, "Effective Information Retrieval Using Similarity Function: Horng and Yeh Coefficient" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 8, August 2013
17. Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert, “A Survey of Binary Similarity and Distance Measures” International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 11, November 2013.
18. Sung-Hyuk Cha, “Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions” International Journal of Mathematical Models and Methods in Applied Sciences, Issue 4, Volume 1, 2007.
19. Md. Abu Kausar, Md. Nasar, Sanjeev Kumar Singh, “A Detailed Study on Information Retrieval using Genetic Algorithm” Journal of Industrial and Intelligent Information Vol. 1, No. 3, September 2013.
20. Hazra Imran, Aditi Sharan, "A Framework for Efficient Document Ranking Using Order And Non-order Based Fitness Function" Proceedings of the International Multiconferrence of Engineers and Computer Scientists, Vol. 1, March 17-19 2010.