Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Effective Information Retrieval Using Similarity Function: Horng and Yeh Coefficient

Manoj Chahal*
Master of Technology (Dept. Of Computer Science and Engineering) GJUS&T, Hisar, Haryana, India

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

Abstract – As the information in the web world is growing rapidly. It is difficult to retrieve relevant information. In this context search engines has become a valuable tool for user to retrieve relevant information. Search engine use Information Retrieval System and Genetic Algorithm to retrieve relevant information. Finding relevant information according to user's need is still a challenge. In this paper cosine and Horng & Yeh similarity Function is used to increase the efficiency of Information Retrieval System. Horng and Yeh similarity Function is applied using Genetic Algorithm to retrieve relevant information.

Keywords: Information Retrieval, Vector Space Model, Database, Similarity Measure, Genetic Algorithm.

I. INTRODUCTION

There is a large amount of information stored in internet or web world. It may consist of documents, video, audio, images files etc. To extract that information IRS (Information Retrieval System), Genetic Algorithm and Matching Function are used. Genetic Algorithm and Matching Function is use to improve relevance of retrieved document which is relevant to the user query.

A. Information Retrieval

Information retrieval is a process in which document database is searching for documents which is relevant to user query. Web search engine is one of the examples of IRS (Information Retrieval System). In IRS user, puts query then that query is matched with the documents in database with the help of matching function and relevant document is extracted and given as output to user.

Components of Information Retrieval System

There are three basic component of Information retrieval system. They are Documentary Database, Query Subsystem and matching function.

1) *Query subsystem:*-Query subsystem is a system which allow user to formulate their queries and present the relevant documents retrieved by the system for user's query.
2) *Matching function:* - Matching function compares both query and documents in database and gives a value which measures the similarity between query and documents. With the help of this value, relevant documents from database are retrieved.
3) *Document database:* - It is the storage space where all the documents are stored. Along with documents it also represents their information content. Matching Function compare all the documents of document database with the user query and extract relevant document from database.

B. Similarity Measures

Similarity Measures is a function use to measure the degree of similarity between query and documents. It measures how much the query and document is similar with each other. It gives a value which decides the degree of similarity. In order to find the similarity first query and document are converted into vector form.

Some of important similarity measures are:

1) Cosine similarity measure

It is one of the most important similarity functions. It is used to measure the angle between query and document. Lower the angle higher the degree of similarity and higher the angle lower the degree of similarity between query and document.

Cosine formulation as shown below:

$$\cos\theta = \frac{\sum_{i=1}^{t} x_i * y_i}{\sqrt{\sum_{i=1}^{t} x_i^2}\sqrt{\sum_{i=1}^{t} y_i^2}}$$

Where x and y are query and document vectors.

2) Horng and Yeh similarity measure

Horng & Yeh formulation as shown below:

$$F = \frac{1}{|D|}\sum_{i-1}^{|D|}\left(r(di)\sum_{j-1}^{|D|}\frac{1}{j}\right)$$

Where D is total of document retrieval and r (d) is a function of relevant document d is set to 1 if a keyword present in document otherwise it is to 0. The formulation of Horng and Yeh represented the value of similarity measure in Genetic Algorithm. The fitness with a higher score reflects a higher probability similarity of document. Application of this technique will facilitate searching and retrieval of the required document from one or more databases based on the similarity level

C. Vector Space Model

Vector Space Model is one of the Information Retrieval System model .In this model both Query and document are represented in vector form. Then with the help of similarity measure degree of similarity between query and document is calculated. Document is retrieve from database based on the degree of similarity between query and document. This means that document with higher degree of similarity should be retrieved by IRS and place in higher position in the list of retrieved documents.

II. Previous Works on Information Retrieval

There are several studies that used genetic algorithm in information retrieval system to optimize the user query.

Nor Hashimah Sulaiman and Daud Mohamad [1] described a similarity measure for soft set based on jaccard similarity coefficient. Two considerations were proposed, first similarity due to the compared parameter and second similarity between value set and parameters. Mahesh A. Sale et al. [2] described a system to extract information from table in web pages. The system transforms the information to form that a computer can understand. Vaibhav Chaudhary, Pushpa Rani Suri [3] discussed the impact of optimization using genetic algorithm and share genetic algorithm on multimodal image registration by considering mutual information concept. P.Pradeep Kumar, Naini.Shekhar Reddy et al. [4] described a method to measure the semantic similarity between words in web world. It was proposed that a lexical pattern extraction algorithm for finding semantic relation between words. P.Iswarya, V.Radha [5] discussed cross language text retrieval process. Different translations and their various approaches with merits and demerits are described and cross language text retrieval in different languages in different domain are compared. Pragati Bhatnagar et al. [6] discussed the applications of GA for improving retrieval efficiency of IRS. GA was used to find an optimal set of weights for components of combined similarity measure consisting of different standard similarity measures that are used for ranking the documents. E man Al Mashagba, Feras Al Mashagba et al. [7] described Genetic Algorithm to find optimal solution for difficult problem. Different similarity measures in vector space model are used and for each similarity measure different GA using different crossover and mutation technique are compared. Gokul Patil and Amit Patil [8] discussed text mining to find useful information from web database. Algorithm to extract data on web and appointed website according to user's request are described.

Mohammad Othman Nassar, Feras Al Mashagba et al. [9] described optimization technique to optimize user query in Arabic data. To optimize query they used GA with different fitness, different crossover and mutation technique.All this technique in Boolean model are applied. S.Siva Sathya and Philomina Simon [10] described document crawler which is used to extract information from web database. As volume of data in web is large they used GA to extract relevant data. Three main steps to extract data from database were proposed. First extract data using document crawler second applying GA to get relevant data third applying result from GA to IRS to get better result. Siti Nurkhadijah Aishah Ibrahim et al. [11] presented a model of hybrid GA-Particle Swarm Optimization (HGAPSO) based query optimization for Web information retrieval. The keywords are used to produce new keywords that are related to the user search. Anna Huang [12] compared and analyzed the effectiveness of these measures in partitioned clustering for text document datasets. The standard K means algorithm and seven text document datasets and five distance and similarity measures that have been most commonly used in text clustering was used. Chengjun Liu [13] proposed that popular whitened cosine similarity measure is related to the Bayes decision rule under specification assumptions and presented two new similarity measures first PRM whitened cosine similarity measure and second within class whitened cosine similarity measure. M. Zolghadri-jahromi, and M.R. Valizadeh [14] described a query sensitive similarity measure mechanism to measure the similarity of two documents. In the first step identified the sources of information that may be used for this purpose. In the second step it was proposed that a query sensitive similarity

measure based on these information sources. Finally it was proposed that a query sensitive similarity measure parametric that simultaneously makes use of the product and weighted sum to fuse the information from the identified sources. Poltak Sihombing, Abdullah Embong et al. [15] described Horng and Yeh formulation in IRS and compared it with jaccard and dice similarity measure. All the technique is applied using GA in IRS. Zhengyu Zhu, Xinghuan Chen et al. [16] described relevance feedback technique to retrieve relevant information. Genetic Algorithm to optimize user query and retrieve web information are applied. Cristina Lopez Pujalte, Felix de Moya Anegon et al [17] described various order based fitness functions than evaluate efficiency of genetic algorithm using this fitness function for relevance feedback. Philomina Simon and S. Siva Sathya [18] described a general frame work of information retrieval system. The applicability of genetic algorithm was discussed in different areas of information retrieval such as genetic mining, query optimization, document clustering, and query optimization etc. P. Pathak, M. Gordon [19] described method to retrieve relevant information from database three paradigms of research in information retrieval are probabilistic IR, knowledge based IR and artificial intelligence based IR are described. S. Brin and L. Page [20] described the working of search engine and also explain the architecture of search engine.

III. Genetic Algorithm

The genetic algorithm is a probabilistic search algorithm which is used for optimization of difficult problem. It is based on Darwinian principle of natural selection. It exploits and explores the document search space. The basic operators used by genetic algorithm are selection, crossover and mutation. By using these operators complex problems can be easily solved.
Genetic Algorithm basic components are:

A. Chromosome Representation

Chromosomes are the initial input given to GA. All the documents and query are first converted into chromosome. This is given as input to the genetic algorithm.

B. Fitness Function

Fitness Function gives a value which is used to calculate the similarity between query and document. Based on this value chromosome is selected for selection mechanism.

C. Selection operator

Selection is the process in which chromosomes is selected for next step or generation in genetic algorithm based on fitness value of chromosomes. Poor chromosome or lowest fitness chromosome selected few or not at all.

D. Crossover operator

Crossover is one of the basic operators of Genetic algorithm. The performance of GA depends on them. In crossover two or more parent chromosomes is selected and a pair of genes are interchanging with each other.

E. Mutation operator

Mutation is a process in which gene of the chromosome is changed. In one point mutation if gene is 0 then change it into 1 and if gene is 1 then change it into 0.

IV. Process of Experiment

- Extract all the words from documents.
- Remove stop words.
- Both query and documents are encoded into chromosomes.
- Encoded chromosomes are given as input to Genetic Algorithm.
- Run Genetic Algorithm until stopping criteria met.

V. Experiment

In this section we discuss about how experiment is conducted and result occur during experiment. In our experiment we put query in search engine and extracted top 10 documents. Encoding documents into chromosome which gives the initial population for input to Genetic algorithm and test Horng and Yeh fitness function with set of parameters: Probability of Crossover (P_c) and Probability of mutation (P_m) to compare the efficiency of Information Retrieval System. In our experiment we also study the effect of different value of crossover and mutation on chromosome. We use crossover probability (P_c = 0.7, 0.8, 0.9) and mutation probability (P_m=0.1, 0.05, 0.01).

VI. Result

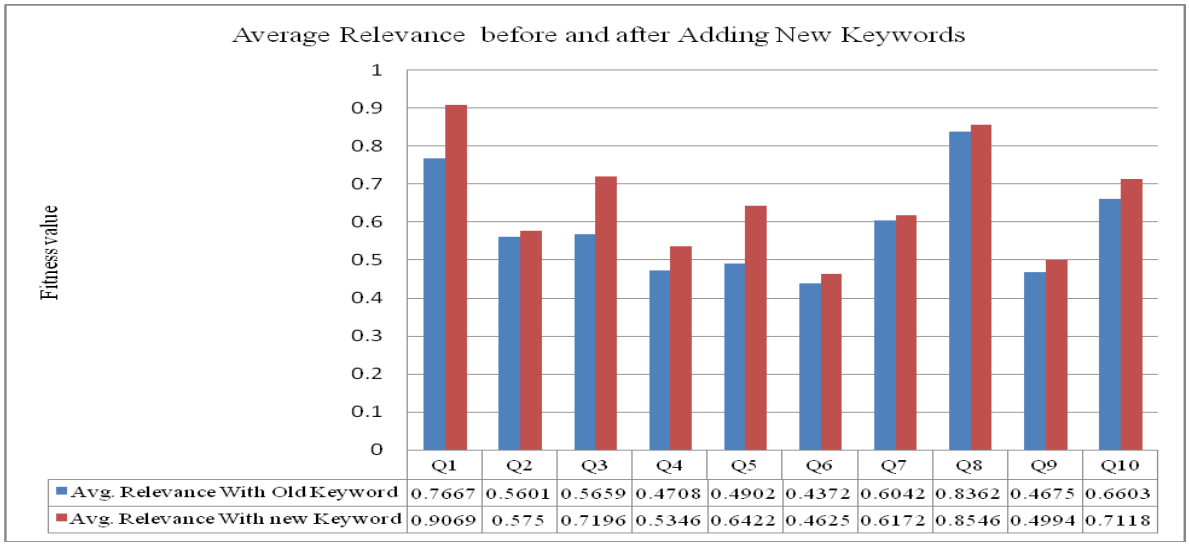
A. Adding new Keyword and Calculating Percentage of Improvement:-

P_c=0.5 p_m=0.005

Table 1.1: Percentage Improvement in Average Relevance after Adding New Keyword.

Query	Avg. Relevance With Old Keyword	Avg. Relevance With new Keyword	% Improvement
Q1	0.7667	0.9069	15.4598%
Q2	0.5601	0.5750	2.5913%
Q3	0.5659	0.7196	21.35%

Q4	0.4708	0.5346	11.934%
Q5	0.4902	0.6422	23.66%
Q6	0.4372	0.4625	5.4702%
Q7	0.6042	0.6172	2.0511%
Q8	0.8362	0.8546	2.152%
Q9	0.4675	0.4994	6.3906%
Q10	0.6603	0.7118	7.2337%



Graph 1.1 Average relevance before and after adding new keywords.

Table 1.1 shows the average relevance with new keyword and old keyword and also show the percentage improvement after adding new keyword. Graph 1.1 shows graphical representation for avg. relevance with old and new keywords.

B. Applied different value of Crossover and Mutation:-

P_c =0.7

Table 1.2: Generation and Average Convergence Value at Different Mutation Probability (P_m =0.1, 0.05, 0.01) at P_c = 0.7.

Query	P_M =0.1		P_M =0.05		P_M =0.01	
	Generation	Conv. On Avg. Value	Generation	Conv. On Avg. Value	Generation	Conv. On Avg. Value
Q1	1000	0.8584	1000	0.9887	33	1
Q2	1000	0.9094	1000	0.9799	19	1
Q3	1000	0.8570	1000	0.9874	43	1
Q4	1000	0.9036	1000	0.9876	21	1
Q5	1000	0.8438	1000	0.9645	16	1
Q6	1000	0.8568	1000	0.9688	22	1
Q7	1000	0.8914	1000	0.9905	15	1
Q8	1000	0.9352	1000	0.9887	20	1
Q9	1000	0.8719	1000	0.8936	26	1
Q10	1000	0.9449	1000	0.9874	28	1

P_c =0.8

Table 1.3: Generation and Average Convergence Value at Different Mutation Probability (P_m =0.1, 0.05, 0.01) at P_c = 0.8.

Query	P_M =0.1		P_M =0.05		P_M =0.01	
	Generation	Conv. On Avg. Value	Generation	Conv. On Avg. Value	Generation	Conv. On Avg. Value
Q1	1000	0.9213	1000	0.9656	22	1
Q2	1000	0.8852	1000	0.9898	26	1
Q3	1000	0.7547	1000	0.9874	33	1
Q4	1000	0.9123	1000	0.9605	36	1
Q5	1000	0.8862	1000	0.9267	14	1
Q6	1000	0.8413	1000	0.9847	26	1
Q7	1000	0.9066	1000	0.9832	14	1
Q8	1000	0.8025	1000	0.9774	24	1
Q9	1000	0.8856	1000	0.9543	21	1
Q10	1000	0.9827	1000	0.9774	18	1

P_c =0.9

Table 1.4: Generation and Average Convergence Value at Different Mutation Probability (P_m =0.1, 0.05, 0.01) at P_c = 0.9.

Query	P_M =0.1		P_M =0.05		P_M =0.01	
	Generation	Conv. On Avg. Value	Generation	Conv. On Avg. Value	Generation	Conv. On Avg. Value
Q1	1000	0.9457	1000	0.9854	18	1
Q2	1000	0.9577	1000	0.9949	16	1
Q3	1000	0.7785	1000	0.9980	22	1
Q4	1000	0.9209	1000	0.9699	34	1
Q5	1000	0.9566	1000	0.9847	13	1
Q6	1000	0.9147	1000	0.9918	19	1
Q7	1000	0.9127	1000	0.9905	13	1
Q8	1000	0.9574	1000	0.9843	21	1
Q9	1000	0.9587	1000	0.9742	17	1
Q10	1000	0.9489	1000	0.9898	14	1

Tables 1.2, 1.3 and 1.4 show the effect of mutation and crossover over the chromosome. According to tables mutation probability P_m=0.1, (P_c=0.7, 0.8, 0.9) no query converge at one chromosome. At P_m=0.05 no convergence take place but at P_m=0.01 all chromosomes converge at one in less number of generation. All this shows that less mutation rate is best for these queries.

VII. **Conclusion and Future Works**

There is a vast amount of information in web world. It is difficult to retrieve relevant information as per user requirement. To retrieve relevant information Genetic Algorithm, Information retrieval System and Similarity measure is used. Genetic Algorithm and Horng and Yeh similarity function is used to measure the similarity between query and documents. Horng and Yeh similarity function, vector space model and Genetic Algorithm are applied to increase the efficiency of relevant information retrieval. It is observed that average relevance of documents increases by applying Horng and Yeh formulation in GA. It means Horng and Yeh have refined our search space. It is also shown that if mutation rate is less then all chromosome converge at one in less number of generation and less mutation is best for queries. Average relevance of document can be increased by applying other methods. In this paper cosine and Horng & Yeh formulation is applied but this work can also be done by applying other similarity measure e.g. jaccard, dice etc with Horng and Yeh formulation and compare the result with each other. In this paper binary vector is applied but this work can also be done with weighted vector.

References

[1] Nor Hashimah Sulaiman and Daud Mohamad, "A jaccard based similarity measure for soft sets", *IEEE Symposium on Humanities, Science and Engineering Research*, pp.659-663, 2012.

[2] Mahesh A. Sale, Pramila M. Chawan, Prithviraj M. Chauhan, "Information extraction from web tables", *International Journal of Engineering Research and Application*, vol. 2, no. 3, pp. 313-318, Jun 2012.

[3] Vaibhav Chaudhary, Pushpa Rani Suri, "Genetic algorithm v/s share genetic algorithm with roulette wheel selection method for registration of multimodal images", *International Journal of Engineering Research and Application*, vol. 2, no. 4, pp.365-370, Aug. 2012.

[4] P.Pradeep Kumar, Naini.Shekhar Reddy, R.Sai Krishna et al., "Measuring of semantic similarity between words using web search engine approach", *International Journal of Engineering Research and Application*, vol. 2, no. 1, pp. 401-404, Feb. 2012 .

[5] P.Iswarya, V.Radha, "Cross language text retrieval", *International Journal of Engineering Research and Application*, vol. 2, no. 5, pp. 1036-1043, Oct. 2012.

[6] Pragati Bhatnagar and 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*, ISSN 2250-2459, vol. 2, no. 6,pp. 249-256, Jun. 2012.

[7] E man Al Mashagba , Feras Al Mashagba and Mohammad Othman Nassar, "Query optimization using genetic algorithm in the vector space model", *International Journal of Computer Science*, ISSN 0814-1694, vol. 8, no. 3, pp. 450-457, Sept. 2011.

[8] Gokul Patil, Amit Patil, "Web information extraction and classification using vector space model algorithm", *International Journal of Emerging Technology and Advanced Engineering*, ISSN 2250-2459, vol. 1, no. 2, pp. 70-73, Dec. 2011.

[9] Mohammad Othman Nassar, Feras Al Mashagba and Eman Al Mashagba, "Improving the user query for the boolean model using genetic algorithm", *International Journal of Computer Science*, vol. 8, no. 1, pp. 66-70, Sept. 2011.

[10] S.Siva Sathya and Philomina Simon, "A document retrieval system with combination terms using genetic algorithm", *International Journal of Computer and Electrical Engineering*, vol. 2, no. 1, pp.1-6, Feb. 2010.

[11] Nurkhadijah Aishah Ibrahim, Ali Selamat, Mohd Hafiz Selamat, "Query optimization in relevance feedback using hybrid GA-PSO for effective web information retrieval", *IEEE Transaction* DOI 10.1109, pp. 91-96, 2009.

[12] Anna Huang, "Similarity Measures for Text Document Clustering", *Proceedings of the New Zealand Computer Science Research Student Conference*, 2008.

[13] Chengjun Liu, "The bayes decision rule induced similarity measures", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1086-1090, 2007.

[14] M. Zolghadri-jahromi, and M.R. Valizadeh, "A proposed query-sensitive similarity measure for information retrieval", *Iranian Journal of Science & Technology*, Shiraz University, vol. 30, no. B2, pp.171-180, 2006.

[15] Poltak Sihombing, Abdullah Embong, Putra Sumari, "Comparison of document similarity in information retrieval system by different formulation", *Proceedings of 2nd IMT-GT Regional Conference on Mathematics Statics and Application*, Malaysia, Jun. 2006.

[16] Zhengyu Zhu, Xinghuan Chen, Qihong Xie, Qingsheng Zhu, "A GA based query optimization for web information retrieval", *International Conference on Intelligent Computing*, pp. 2069-2078, Aug. 2005.

[17] Vicente P., Cristina P., "Order-Based Fitness Functions for Genetic Algorithms Applied to Relevance Feedback", *Journal Of The American Society For Information Science And Technology*, 54(2):152–160, 2003.

[18] P. Simon, and S.S. Sathya, "Genetic algorithm for information retrieval", *International Conference on Intelligent Agent & Multi-Agent Systems (IAMA)*, ISBN: 978-1-4244-4710-7, pp. 1 – 6, 2009.

[19] P. Pathak, M. Gordon and W. Fan. "Effective information retrieval using genetic algorithms based matching functions adaption", in: Proc. *33rd Hawaii International Conference on Science (HICS), Hawaii*, USA, 2000.

[20] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proc. Seventh Int'l Conf. World Wide Web* (WWW '98), pp. 107-117, 1998.

