Information Retrieval using Jaccard Similarity Coefficient

Manoj Chahal*

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

Abstract – *Similarity measure define similarity between two or more documents. The retrieved documents are ranked based on the similarity of content of document to the user query. Jaccard similarity coefficient measure the degree of similarity between the retrieved documents. In this paper we retrieved information with the help of Jaccard similarity coefficient and analysis that information. All this is performed with the help of Genetic Algorithm. Due to exploring and exploiting nature of Genetic Algorithm it gives optimal result of our search. Genetic algorithm use Jaccard similarity coefficient to calculate similarity between documents. Value of jaccard similarity function lies between 0 &1 .it show the probability of similarity between the documents.*

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

I. INTRODUCTION

A search engine is a tool that allows people to find information on the Internet. Information may consist of web pages, images, information and other type of files. Some search also mine data available in news, books, database, or open directories. To retrieve relevant information search engine use Information Retrieval System.

1 *Information Retrieval System*

The various parts of information retrieval is

- user
- Query Subsystem
- Matching Mechanism
- Document Database

User:-

User is a person who put the request on the information retrieval system on the bases of this request information is retrieved from the database.

Query subsystem:-

Query subsystem is a system which formulate user request into query. It contains a query language that collects the rules to generate query.

Matching function:-

Matching function compare the similarity between the query and document in the database. Based on this document are retrieved.

Document database:-

This component stores the documents and the representations of their information contents. It is associated with the indexer module which automatically generates are presentation for each document by extracting the document contents.

There are three information retrieval models have been studied and developed in the information retrieval area are

2. Similarity Measures:-

Similarity Measures is a function which is used to measure the similarity between user query and documents. User put query on the search engine than with the help this user gets the relevant information from the web world.

- Jaccard similarity measure

It measures similarity between the two documents. The value is between 0 and 1. 0 show that documents are dissimilar and 1 show those documents are identical with each other. Value between 0 and 1 show the probability of similarity between the documents.

Jaccard formulation as shown below:

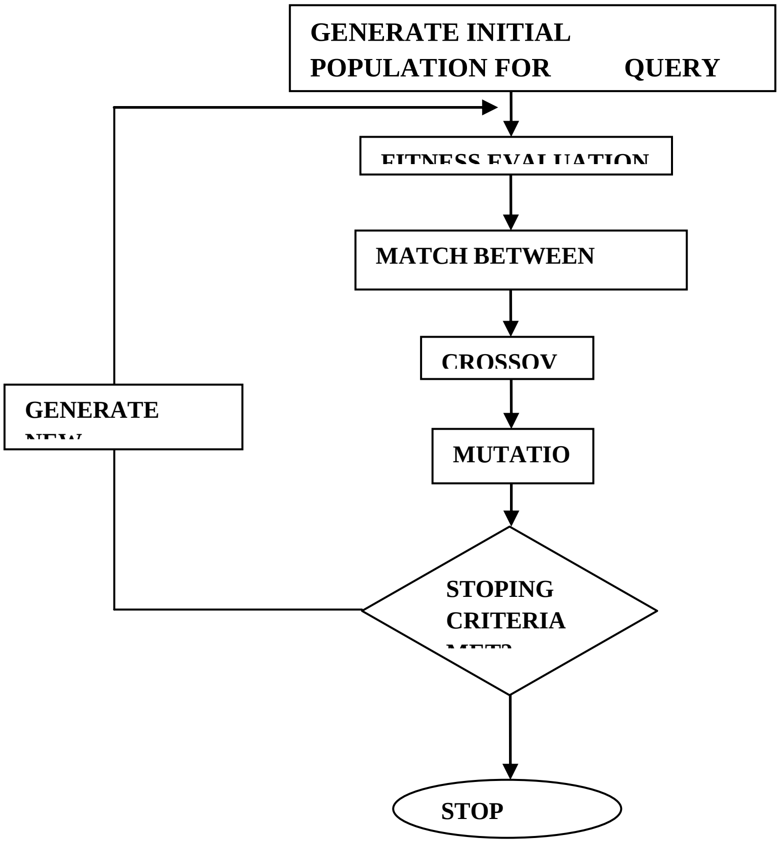
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

II. GENETIC ALGORITHM

Genetic algorithm is an algorithm used to get the optimal solution for a given problem. It is used to exploits and explores the search space of problem in order to get the optimal solution. It is based on the Darwinian principle of natural selection.

The components of genetic algorithm are:-

- Selection
- Mutation
- Crossover



Flow Chart of Genetic algorithm [10]

III PREVIOUS WORKS ON INFORMATION RETRIEVAL

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

E men Al Mashagba et al [1] described various different similarity measures like dice, cosine, Jaccard etc in vector space model and compare each similarity measures using genetic algorithms approach based on different fitness functions, different mutations and different crossover to find the best solution of the given query. Mohammad Othman Nassaret al [2] described binary model using genetic algorithm with different fitness function and different mutation strategy to retrieving relevant information and query optimization. .Pradeep Kumar, Naini.Shekhar Reddy, R.Sai Krishna et al[3] described semantic similarity between words and to measure semantic similarity they applied lexical pattern extraction algorithm and sequential pattern clustering algorithm.

Poltak Sihombing et al [4] described Information retrieval system using genetic algorithm and various matching functions to compare the similarity between the user query and document database. Gokul Patil and Amit Patil[5] described web based text mining problem and step to solve that problem and filter out just those that have the desired meaning. J.Allaan, Jay Aslam et al. [6] described various area of information retrieval system and also describes major challenges within each of those areas. Seung-Seok Choi, Charles C. Tappert [7] described various similarity and distance measures. Each of them is differently defined by its own synthetic properties. Some include negative matches and some do not. Some use simple count difference and some utilize complicated correlatio Pragati Bhatnagar [8] 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. Vaibhav Chaudhary, Pushpa Rani Suri [9] discussed the impact of optimization using genetic algorithm and share genetic algorithm on multimodal image registration by considering mutual information concept.

IV. EXPERIMENT

Step to conduct experiment

- First user search the required documents
- Then convert this documents into initial chromosome as input to Genetic Algorithm
- Applying Jaccard similarity function
- Then apply selection, crossover and Mutation

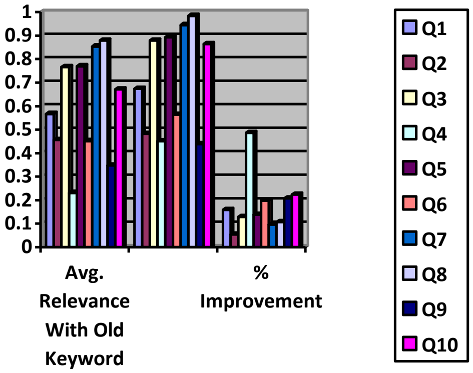
- Genetic algorithm run until stopping criteria met

V. RESULT

Adding new Keyword and Calculating Percentage of Improvement :

Query	Avg. Relevance With Old Keyword	Avg. Relevance With new Keyword	% Improvement
Q1	0.5673	0.6745	15.89%
Q2	0.4563	0.4823	5.39%
Q3	0.7659	0.8790	12.86%
Q4	0.2317	0.4512	48.64%
Q5	0.7698	0.8934	13.83%
Q6	0.4517	0.5634	19.82%
Q7	0.8540	0.9452	9.64%
Q8	0.8792	0.9845	10.69%
Q9	0.3476	0.4387	20.76%
Q10	0.6709	0.8645	22.39%

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



VI. CONCLUSION AND FUTURE WORKS

It is observed that average relevance of documents increases by applying Jaccard Similarity Function in GA. It means Jaccard Similarity Function explore and exploit our search space. Average relevance of document can be increased by applying other methods. In this paper Jaccard Similarity Function is applied but this work can also be done by applying other similarity measure 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] 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.

[2] 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.

[3] 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 .

[4] 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.

[5] 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.

[6] J.Allaan, Jay Aslam et al. “Challenges in Information Retrieval and Language Modeling “ , *Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002.*

[7] Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert,” A Survey of Binary Similarity and Distance Measures”,. *Department of computer science, Pace University*

[8] 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, June 2012*

[9] Vaibhav Chaudhary, Dr. 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, August 2012.*

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