www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issues 6 June 2016, Page No. 17070-17075

Effect on Information Retrieval Using One and Two Point Crossover

Manoj Chahal [*]

Master of Technology (Dept. Of Computer Science and Engineering) GJUS&T, Hisar, Haryana
e-mail : manojchahal008@gmail.com

Abstract- Information in digital world growing at very high speed. It is impossible to retrieve relevant and important information. To retrieve relevant information search engine is used. Genetic Algorithm and Information Retrieval System is used by search engine in order to retrieve relevant information. But retrieving information as user requirement is still difficult. In this paper information is retrieve using Genetic algorithm with one and two point crossover and cosine and Horng & Yeh similarity Function is used as fitness function in Genetic Algorithm.

Keywords: Genetic Algorithm, One point crossover, Two point crossover, Information Retrieval, Vector Space Model, Database, Similarity Measure.

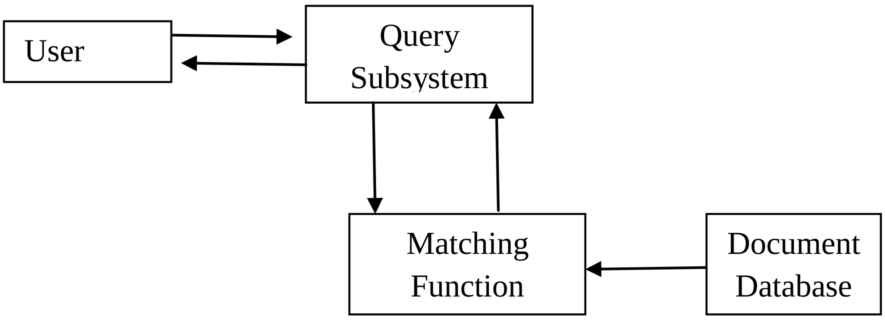
INTRODUCTION

Large quantity of information available in internet world .this information is available in many forms. In order to retrieve relevant information IRS (Information Retrieval System) is used. Information Retrieval System use matching function, query subsystem to retrieve required information from database.

1 Information Retrieval Architecture

The various parts of information retrieval architecture is

- user
- Query Subsystem
- Matching Mechanism
- Document Database



Information retrieval begin when user enter query that query process in sub query system. Then matching function match the required documents with query and data from database is retrieved and given to the user as the result of their query.

2 Information Retrieval Models

The ultimate goal of information retrieval system is satisfy user information requirement .To retrieve information there is three information retrieval models they are

- Boolean model
- Vector space model
- Probabilistic model

Vector space model

It is one of the information retrieval models. In this model a document is viewed as a vector in n-dimensional document space and each term represents one dimension in the document space. Document retrieval is based on the measurement of similarity between the query and document.

II. **GENETIC ALGORITHM**

Genetic algorithm is an adaptive search algorithm which is based on Darwinian principle of natural selection and genetic. It is used to optimization of difficult problem. it explore and exploit the search space to find the optimal solution. It follows the principles laid down by Charles Darwin of survival of the fittest.

There are three basic operators and parameters used in genetic algorithm are

- Selection

  Selection is the process in which chromosome is selected for next step in Genetic Algorithm. Poor chromosome or lowest fitness value chromosome selected few or not at all.

- Crossover

  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.

- Mutation

  Mutation is the occasional introduction of new features in to the solution strings of the population pool to maintain diversity in the population.

III. **PREVIOUS WORKS ON INFORMATION RETRIEVAL**

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

Zhengyu Zhu, Xinghuan Chen et al. [1] described vector model which is based on similarity measurement. Which extract documents having high similarity for the giving query are retrieved first. Each query and document is represented by chromosomes. This chromosome is feed into genetic algorithm process. After completing the whole process it gives an optimized query chromosome for information retrieval. Sergey Brin and Lawrence[2] described Page crawler, page rank, indexer etc which are used to retrieve useful information. Crawlers are small application program which used to collect information from web. With the help of crawlers search engine database created. Page rank is used to give an order to the web page according to the user query. With the help of this tools search engine give useful information to the user. Gokul Patil and Amit Patil[3] described step to extract information in vector information retrieval model . it also described how Web based text mining effort that collect a significant number of Web mentions of a subject.

Poltak Sihombing, Abdullah Embong, Putra Sumari[4] described comparison of document similarity by using different matching function. Cristina Lopez Pujalte, Felix de Moya Anegon et al [5] described various order based fitness functions than evaluate efficiency of genetic algorithm using this fitness function for relevance feedback. 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. Anna Huang [7] described similarity measure based on clustering technique. It describe that how clustering technique organizes large quantity of unordered text documents into a small number of meaningful and coherent clusters.

Poltak Sihombing, Abdullah Embong et al. [8] described Horng and Yeh formulation in IRS and compared it with jaccard and dice similarity measure. Mahesh A. Sale et al [9] described how information is extracted from web table. Tables are stored in web which is important source of information. In order to extract table's information from web pages table extraction, web page processing, table validation, table normalization, table interpretation and attribute value pair formation are applied. Chahal et al [10] describe information retrieval using Horng and Yeh similarity function and also describe effect of different value of crossover and mutation.

IV. EXPERIMENT

Step to conduct experiment

- First search 10 document with the help of search engine
- Encoding document into chromosome so that initial population is created
- This initial population serve as input to genetic algorithm
- Cosine and Horng and Yeh function used as fitness function
- Appling genetic algorithm in document with one point (GA)and two point crossover(GA1)
- Compare one point and two point cross over with each other
- We use crossover probability (P_c = 0.8) and mutation probability (P_m=0.7).
- Run Genetic Algorithm until stopping criteria met.

V. RESULT

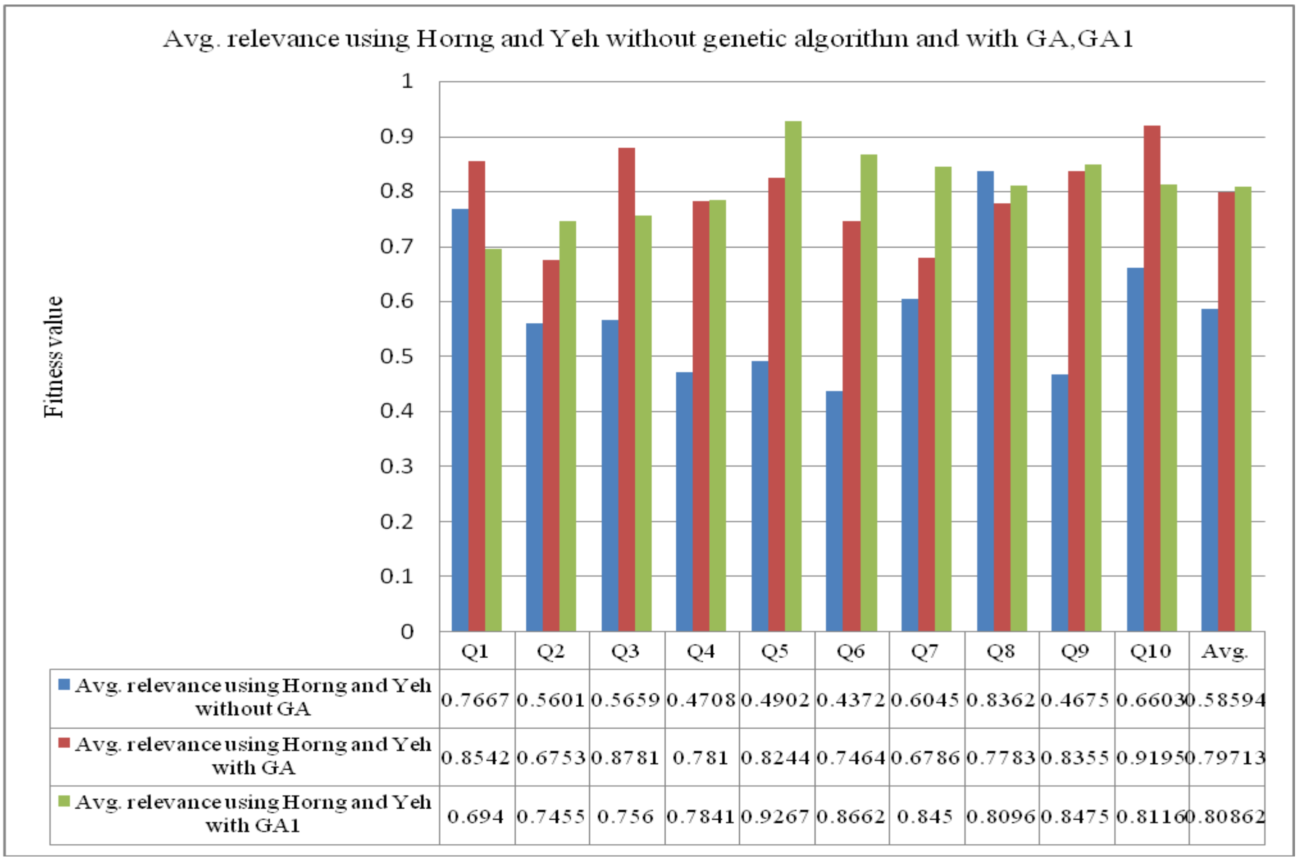
In this experiment, different technique of crossover is applied with the help of GA. Then compare average fitness value with Horng and Yeh Coefficient.

P_c =0.8, P_M =0.7

Query	Generation	Horng and Yeh (Fitness value without GA)	(GA) One point crossover and one point mutation (Avg. Fitness value with GA)	(GA1) two point crossover and one point mutation (Avg. Fitness value with GA1)
Q1	1000	0.7667	0.8542	0.6940
Q2	1000	0.5601	0.6753	0.7455
Q3	1000	0.5659	0.8781	0.7560
Q4	1000	0.4708	0.7810	0.7841
Q5	1000	0.4902	0.8244	0.9267
Q6	1000	0.4372	0.7464	0.8662

Q7	1000	0.6045	0.6786	0.8450
Q8	1000	0.8362	0.7783	0.8096
Q9	1000	0.4675	0.8355	0.8475
Q10	1000	0.6603	0.9195	0.8116
Avg.		0.58594	0.79713	0.80862

Table 1.1: Appling GA, GA1 and Horng and Yeh on queries and calculating average fitness value



Graph 1.1 Avg. relevance using Horng and Yeh without genetic algorithm and With GA, GA1

Graph 1.1 shows that Average relevance using Horng and Yeh without genetic algorithm and with GA and GA1. In without genetic algorithm simple Horng and Yeh are used to calculate relevance genetic algorithm is not applied. In GA and GA1 Horng and Yeh is applied with genetic algorithm. In GA one point crossover and one point mutation is used and in GA1 two point crossover and one point mutation is used. Blue bar show avg. relevance using Horng and Yeh without GA, red and green bar show Avg. relevance using Horng and Yeh with GA and GA1. Y axis in graph 1.1 shows fitness value and X axis show queries.

Graph 1.1 shows that avg. relevance using Horng and Yeh with GA and GA1 is greater than without genetic algorithm. It also shows that avg. relevance for GA1 as compare to GA is slightly greater. This shows that GA1 is slightly better than GA.

Query	(GA) One point crossover and one point mutation (% improvement)	(GA1) two point crossover and one point mutation (% improvement)

Q1	10.24	-10.47
Q2	18.72	24.86
Q3	35.55	25.14
Q4	39.71	39.95
Q5	40.53	47.10
Q6	41.42	49.52
Q7	10.91	28.46
Q8	-7.4	-3.2
Q9	44.04	44.83
Q10	28.18	18.64
Avg.	26.18	27.53

Table 1.2: GA's Improvement in Horng and Yeh Similarity (GA's Improvement %).

The result for different Genetic Algorithm Strategies using Horng and Yeh Coefficient are shown in Tables 1.1 and 1.2. From table 1.1 and 1.2 notices that GA and GA1 give a high improvement as compare to simple Horng and Yeh coefficient without GA.

Tables 1.2 show that GA1 that use two point crossover and point mutation gives slightly improvement than GA which use one point crossover and point mutation and simple Horng and Yeh coefficient.

VI. **CONCLUSION AND FUTURE WORKS**

It is observed that Horng & Yeh formulation with one point and two point crossovers with GA give better result as compare to simple Horng & Yeh formulation without GA and also two point crossover give slightly better performance than one point crossover with GA . Average relevance of document can be increased by applying other methods. In this paper genetic algorithm with one point and two point crossover with Horng & Yeh formulation is applied but this work can be done by applying other similarity measure in place of Horng and Yeh formulation.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[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] 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.

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

[8] Philomina Simon, “A Two Stage approach to Document Retrieval using Genetic Algorithm”, *International Journal of Recent Trends in Engineering*, vol.1, no 1, 526-528, 2009.

[9] 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.

[10] Manoj Chahal and Jaswinder Singh ,” Effective Information Retrieval Using Similarity Function: Horng and Yeh Coefficient”, *International Journal of Advanced Research in computer science and software Engineering* , vol 3 Issue 8 ,pp- 401-406 ,Aug 2013.