© UNIVERSAL RESEARCH REPORTS | REFEREED | PEER REVIEWED
ISSN : 2348 - 5612 | Volume : 05 , Issue : 05 | April – June 2018

Importance of Genetic Algorithm and Similarity Function in Information Retrieval

Manoj Chahal[1] Master of Technology (CSE) G J U Hisar, Haryana, India[1]

Abstract – In every field information plays an important role .To collect important and relevant information in digital world is a challenging task. In this context various searching algorithms and similarity function is used to retrieve relevant information. In this paper genetic algorithm and various similarity functions are discussed. Genetic algorithm and similarity functions help to get the relevant information from the digital world. The objective of this paper is to summarize the whole process and looking into some of the known genetic algorithms and similarity functions to retrieve relevant information.

Keywords: Genetic Algorithm, crossover, Mutation, Similarity Functions, Information Retrieval.

ISSN : 2348-5612 © URR
9 770234 856124

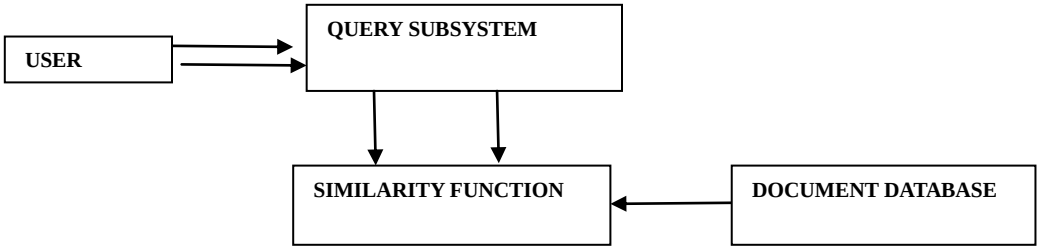
I. INTRODUCTION

Information retrieval is a system which is used to retrieve relevant or important information from various databases or the Internet. The main objective of information retrieval is to help user to find important information quickly and accurately from the Internet or various databases . Genetic algorithm and similarity functions help to achieve this. In information Retrieval System queries are matched with database information using similarity function and genetic algorithm is used to get the optimal value from the databases.

Information Retrieval Architecture

The various parts of Information retrieval architecture is

(1) User
(2) Query Subsystem
(3) Similarity Function
(4) 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 allow user to formulate their queries and present the relevant documents retrieved by the system for user's query.

Similarity function: Similarity function compares both query and documents in database and give a value which measure the similarity between query and documents. With the help of this value, relevant documents from database are retrieved.

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.

Information Retrieval Models:

There are three categories of information retrieval models:-

- **Boolean model**
- **Vector space model**
- **Probabilistic model**

Boolean model: In the Boolean retrieval model the indexer module perform a binary indexing in the sense that a term in a document representation is either significant or not. User queries in this model are expressed using a query language that is based on these term and allows combination of simple user requirement with the logical operator AND, OR and NOT.

Vector space model: 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.

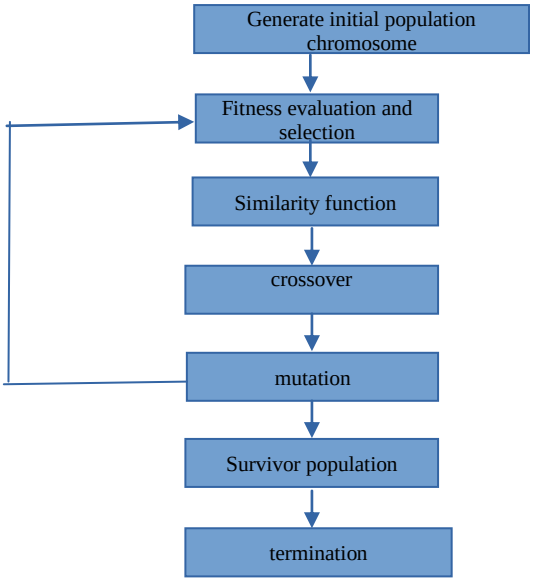
Probabilistic model: This model tries to use the probability theory to build the search function and its operation mode. The information used to compose the search function is obtained from the distribution of the index term throughout the collection of document.

Genetic algorithm:

Genetic Algorithm is a search-based optimization technique based on the principles of Genetics and Natural Selection. It is frequently used to find optimal solutions to difficult problems . It is frequently used to solve optimization problems.

Optimization means for a given input value finding the best possible output value

The genetic algorithm is illustrated in Flow chart.



Generate Initial Population Chromosome: Initially document is converted into chromosome for giving input to Genetic algorithm. These initial chromosomes make initial population for Genetic algorithm.

Fitness evaluation and Selection: It is a process in which fitness value of each chromosome is calculated with the help of fitness functions and selection operation is applied on the entire chromosome with the help of roulette wheel. This fitness value helps us to match between query and documents.

Crossover: It is basic operation of GA. In this operation two parent chromosomes interchange there group of genes and produce child chromosome. Crossover operator is responsible to search different area of search space.

Mutation: Mutation is a random process where one chromosome of a gene is replaced by another to produce a new genetic structure. In genetic algorithm mutation is randomly applied with low probability and modifies elements in the chromosomes. Mutation operation is responsible to keep diversity in the population.

Survivor population: The survivor selection determines which individual population is keeping in next generation. It is selected based on the fitness value of individual population.

Termination: Genetic algorithm is a stochastic search method it is difficult to formally specify convergence criteria. As the fitness of a population may remain static for a number of generations before a superior individual is found the application of conventional termination criteria becomes problematic. A common practice is to terminate the genetic algorithm after a prespecified number of generations and then test the quality of the best members of the population against the problem definition. If no acceptable solutions are found the genetic algorithm may be restarted or a fresh search initiated.

II. PREVIOUS WORKS ON INFORMATION RETRIEVAL

There are various studies that used genetic algorithm and similarity functions in information retrieval system to get the relevant information.

Philomina Simon and S. Siva Sathya[1] described a general frame work of information retrieval system. The applicability of genetic algorithm was discussed in different areas of information retrieval. Pragati Bhatnagar and N.K. Pareek [2] described combined matching function for improving efficiency of information retrieval system. It uses a genetic algorithm for adapting weight of combined similarity function.Sergey Brin and Lawrence[3] 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. Withthe help of crawlers search engine database created. E man Al Mashagba , Feras Al Mashagba and Mohammad Othman Nassar[4] described vector space model to retrieve information from database or Internet and also compare various similarity function.

Anna Huang[5] described how to apply similarity function on text document clustering it also compared and analyzed the effectiveness of these measures in partitioned clustering for text document datasets. Mahesh A. Sale, Pramila M. Chawan, Prithviraj M. Chauhan[6] described the technique for extracting information from table in web pages. The system transforms that information in to computer understandable form. Vaibhav Chaudhary, Pushpa Rani Suri [7] explained the impact of optimization using genetic algorithm and share genetic algorithm on multi modal image registration by considering Mutual information concept. Nor Hashimah Sulaiman and Daud Mohamad[8] 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. Chahal et al [9] describe information retrieval using Horng and Yeh similarity function and also describe effect of different valueof crossover and mutation.Poltak Sihombing, Abdullah Embong et al. [10] described Horng and Yeh formulation in IRS and compared it with jaccard and dice similarity measure.

III. CONCLUSION

There is a huge amount of information stored in Internet and it is increasing exponentially. To search and retrieve important information as per user requirement is a difficult task. To solve this Genetic Algorithm, Information retrieval System and Similarity measure is used. Genetic Algorithm and similarity function is used to measure the similarity between user query and documents. similarity function, vector space model and Genetic Algorithm are applied to increase the efficiency of relevant information retrieval. Average relevance of documents increase by similarity function in GA. It means similarity function and genetic algorithm refined search space. Search space can also further refined by using different crosser over and mutation method.

REFERENCES

[1]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.

[2]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.

[3]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.

[4]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.

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

[6]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.

[7]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.

[8]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.

[9]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.

[10]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.