An Overview of Genetic Algorithm Based Information Retrieval

Suman[1]
Department of Computer Science,
Kurukshetra University, Kurukshetra, India
e-mail: sumansaini235@gmail.com

Pinki Rani[2]
Department of Computer Science,
Kurukshetra University, Kurukshetra, India
e-mail: sainipinki680@gmail.com

Abstract- As the information grows rapidly, searching relevant and up to date information has become a crucial issue. The amount of information and the pages that are similar to each other are also increasing. Information retrieval is a process or method whereby a prospective user of information is able to transform his need for information into an actual list of documents in storage containing information useful to him. An Information Retrieval System (IRS) Can be defined as a system which interprets the contents of the information items and generate a ranking which reflect relevance and retrieves the information more efficiently. This paper intends the study of genetic algorithm based information retrieval using similarity measures like cosine coefficient, jaccard coefficient, dice coefficient. The genetic algorithm aims to optimize the overall relevance estimate by applying a customized fitness function which will make use of local as well as global factors to have the evaluation function distributed over the search space.

Keywords- *Information Retrieval; Similarity Measure; Genetic Algorithm; Fitness Function; Crossover; Mutation.*

*****

I. INTRODUCTION

An Information Retrieval System consists of a software program that facilitates a user in finding the information as the user needs. IR is to provide the users with the documents that satisfy their information needs. IRS have to extract the keywords from the documents and assign weights for each keyword. The user query is the primary representation of user's abstract information needs. An optimal Information Retrieval System (IRS) is one which is able to retrieve only those documents that are relevant to a user's information needs from the document database, but excluding documents that are irrelevant. The field of Information Retrieval being complex and of high-dimension requires a set of techniques or algorithms which can give an approximate solution that is close to optimal. GA is characterized by higher probability of finding good solutions for large and complex problems and has proved to be effective in IR technique of relevance feedback reweighting document term indexing, query reformulation and fuzzy systems, document indexing, document matching and ranking relevance optimization.

II. INFORMATION RETRIEVAL SYSTEM

Information Retrieval System (IRS) is a system used to store items of information that need to be processed, searched and retrieved corresponding to a user's query. Most IRSs use keywords to retrieve documents. The systems first extract keywords from documents and then assign weights to the keywords by using different approaches. Such a system has two major problems. One is how to extract keywords precisely and the other is how to decide the weight of each keyword.

A. Components of IRS

An IRS consists of basic components: User, Documentary Database, Query Subsystem, and Matching mechanism

- **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.
- **The Documentary Database**: This document database stores document along with the representation of their information content. It is associated with the indexer module which automatically generates a representation of each document by extracting the document contents.
- **The Query Subsystem**: It allows the user to specify their information needs and presents the relevant documents retrieved by the system to them. The efficiency of an IRS system significantly depends upon query formation.
- **The Matching Mechanism**: It evaluates the degree to which documents are relevant to user query giving a retrieval status value (RSV) for each document.

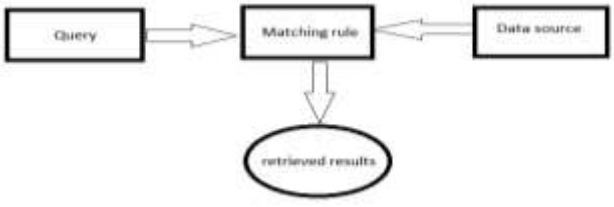

Fig.1 Basic Model of an Information Retrieval System

 ISSN: 2321-8169
652 – 655

III. GENETIC ALGORITHM IN INFORMATION RETRIEVAL

Genetic Algorithms (GAs) are adaptive and best fitted heuristic optimization techniques premised on the evolutionary ideas of natural and genetic selection. They are based on the principles of the evolution via natural selection, generating a population of individuals that undergo selection in the presence of variation-inducing operators, such as mutation and recombination. This process is reiterated a number of times, as it advance towards better and better individuals. GAs are suitable for the information retrieval because of its robustness and quick search capabilities in large and complex search space. GAs specialize in large, complex and poorly understood search spaces where classic tools are inappropriate, inefficient or time consuming. The features of GA are as follows:

- GA has exhibited powerful search capabilities and is equipped with global exploration capabilities, thus suited to Information Retrieval.
- GA possess the property of implicit parallelism, hence can perform search in different regions of document space simultaneously.
- GA is effective in searching through complex, highly nonlinear, multi- dimensional search spaces. They have the ability to manipulate a population of queries rather than a single query.
- GA is highly domain independent and uses probabilistic exploration. It is independent of initial query.

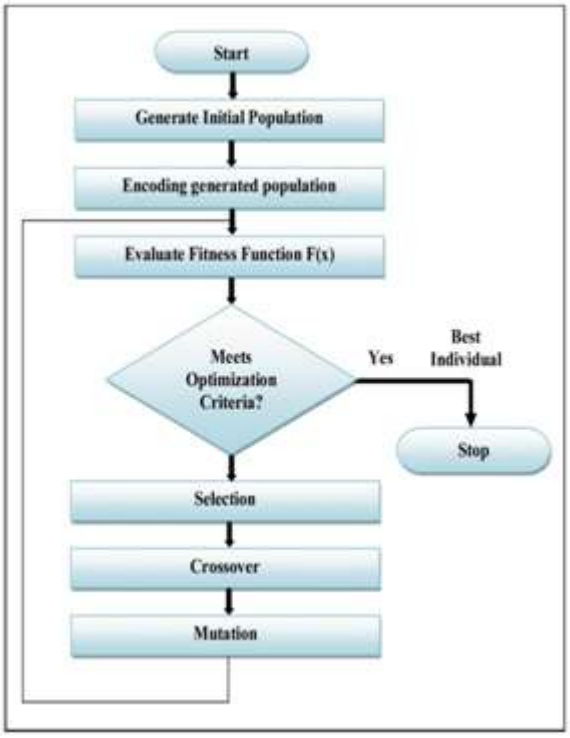

Fig. 2: Generic flowchart of Genetic Algorithm

In Fig. 2, a generic flowchart of Genetic Algorithm has been depicted. The initial population is defined as the collection of individuals that undergo changes to produce new generation. The members of each generation in the Genetic Algorithm process undergo the process of selection, recombination and mutation to form a new population, depending upon their fitness values. The dimension of the GA refers to the dimension of the search space which equals the number of genes in each chromosome. In genetic model, the initial population will consists of the set of documents encoded as genes. The encoding involves representing a document reference number. A fitness value attributed to each chromosome is evaluated at each generation from the current population. The query is sent to the information retrieval system and the query is matched against the chromosomes containing documents as genes for finding the set of more relevant documents. The process of crossover and mutation goes on until a better combination of relevant documents is achieved. The evolutionary process goes on up to a number of generations or the best individual do not change for a number of generations.

A. Chromosome Representation

The representation of chromosomes has profound impact upon the performance of genetic algorithm. These chromosomes use a binary representation, and are converted to a real representation by using a random function. We will have the same number of genes (components) as the query and the feedback documents have terms with non-zero weights. The set of terms contained in these documents and the query is calculated. The size of the chromosomes will be equal to the number of terms of that set, we get the query vector as a binary representation and applying the random function to modify the terms weights to real representation. The chromosome is represented as a set of strings. It uses fixed-length binary strings to represent chromosome, where each position corresponds to a query term.

B. Creation of Initial Population

The initial population creation plays a substantial role in GA process, because the subsequent generations will inherit the characteristics from this generation. Also, the speed at which the optimal solution will be found is dependent upon the quality of the individuals of the first generation. The initial population is a subset of search space and can be created in a number of ways. One of the most popular and widely used method is random selection of individuals, i.e. individuals are selected from the search space without any explicit selection criteria. This search starts with some good individuals but the numbers of documents are limited in the beginning.

C. Fitness Function

The fitness function evaluates the performance of each individual chromosome to judge its contribution in GA. It evaluates the significance or relevance of the document to the user query. Thus selection of an appropriate fitness function is of high importance. The fitness function can either be standard fitness function based on similarity measures or customized fitness function. The similarity function is a measure of degree of closeness of document vector to the user query vector. The most popular similarity measure is cosine similarity measure which is cosine of the angle between document vector and query vector, and yields high results. Other popular similarity measures are Dice coefficient and Jaccard coefficient. These measures were developed specifically for systems implemented using VSM. They are simple, straightforward and their results reflect high performance.

1. Cosine Similarity Measure:

Cosine similarity is most commonly used in high-dimensional positive spaces. In information retrieval and text mining, each term is notionally assigned a different dimension and a document is characterized by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter. The cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$a.b = \|a\|_2 \|b\|_2 \cos\theta$$

Given two vectors of attributes, A and B, the cosine similarity, cos(θ), is represented using a dot product and magnitude as

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

where Ai and Bi are components of vector A and B respectively. In the case of information retrieval, the cosine similarity of two documents will range from 0 to 1, since the term frequencies (tf-idf weights) cannot be negative. The angle between two term frequency vectors cannot be greater than 90°.

2. Dice Similarity Measure:

Dice Similarity measure come under 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. Dice formulation as shown below:

$$Dice_{(A,B)} = \frac{2|A \cap B|}{|A| + |B|}$$

3. Jaccard Similarity Measure:

The **Jaccard index**, also known as **Intersection over** Union and the Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

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

D. Selection Operator

The Selection Mechanism is the most important operator of GA which contributes improving the average quality of the population. The objective of selection operator is to use a selection technique that has low time complexity as well as the potential of selecting healthier parents for crossover operation, while passing a few good chromosomes of the old generation to the next generation, provided that the selected parents do not lead to local optima and do not converge at low performance.

E. Crossover

Crossover is a genetic operator responsible for generating the offsprings from the existing population. The idea behind crossover is that the new chromosome may be better than both of the parents if it takes the best characteristics from each of the parents. Different crossovers have been employed in GA algorithm by researchers.

- ***Single-point crossover*** is the most commonly used crossover in Information Retrieval, as it the most simple crossover type and easy to apply. In single point crossover subparts are exchanged between chromosomes, before or after a randomly selected location. Though it is simple to implement and fast in generating child, it may produce children with lower performance.
- ***2-point*** **and** ***Multipoint crossover*** where two crosspoints or multiple crosspoints are selected for exchange of genes. They provide better diversity than one-point crossover, but reduce speed of crossover and building blocks may disrupt. The multi-point crossover reduces the speed of the

crossover and usually disrupts building blocks, which results in lower performance.

This study aims at selecting a crossover that generate chromosomes with good points by inheriting best features from parents and maintaining performance of the chromosome at the same time.

F. Mutation

Mutation operator introduces a small amount of random perturbation in the chromosome structure, and helps ensure that no point in the search space has a zero probability of being examined. Mutation is traditionally seen as a "background" operator, but research has proven that mutation generally finds better solutions than a crossover-only strategy. There are many goals of applying mutation. They are restoring lost data, exploring variety of data, improving diversity of the solution and reduce the possibility of converging to a local optimum, rather than the global optimum.

G. Termination Criteria

The process of GA ends when either a maximum number of generations have been produced or a satisfactory fitness level has been reached for the population or the population has converged substantially. Widely used termination criteria are:

- Reaching optimal solution (which is often hard, if not impossible, to recognize).
- Processing fixed number of generations.
- Processing certain number of generations without improvement in population.

IV. CONCLUSION

This paper presents the fundamentals of IR and Genetic Algorithm. GA starts with a limited number of individuals from initial population. GA constructs a new generation from old one by following three steps - selection, crossover and mutation. Fitness function evaluates the relevance of information to the user query. The similarity between queries and documents is computed by using any of similarity measures like Cosine coefficient, Dice coefficient, Jaccard coefficient. The retrieved information is sorted by its similarity value with the query.

REFERENCES

[1] H. Chen, “Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms”. Journal of the American Society for Information Science, 46(3), 1995, pp. 194–216.

[2] J. Savoy and D. Vrajitoru, “Evaluation of learning schemes used in information retrieval (CR-I-95-02)”. Universite de Neuchatel, Faculte de droit et des Sciences Economiques, 1996.

[3] M. Gordon, “Probabilistic and genetic algorithms in document retrieval”. Commu- nications of the ACM, 31(10), 1988, pp. 1208–1218.

[4] 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.

[5] M. Gordon. “Probabilistic and genetic algorithms for document retrieval”, Communications of the ACM 31 (10), 1988, pp. 1208–1218.

[6] W. Fan, M.D. Gordon and P. Pathak. “Discovery of context-specific ranking functions for effective information retrieval using genetic programming”, *IEEE Transactions on knowledge and Data Engineering, in press.*

[7] M.P. Smith, M. Smith. “The use of genetic programming to build Boolean queries for text retrieval through relevance feedback”, Journal of Information Science 23 (6), 1997, pp. 423–431.

[8] T. Noreault, M. McGill and M. B. Koll. “A performance evaluation of similarity measures, document term weighting schemes and representation in a Boolean environment”. Information retrieval research. London: Butterworths, 1981.

[9] 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.

[10] 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.

[11] 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.

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