International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.2, Issue No.10, pp : 943-946 1 Oct. 2013

Document Retrieval System using Genetic Algorithm

Mr. A. Kalayanasaravan, Dr. M. Thangamani, Dr. E. T. Venkatesh
Department of Computer Science and Engineering,
Kongu Engineering College, Perundurai, Erode-638 052, Tamilnadu, India,
Email: manithangamani2@gmail.com, kalyanasaravanan@gmail.com, etvkec@gmail.com

Abstract

As information has been increasing enormously in the world, it is very difficult to retrieve information as per the user satisfaction. The main objective of this project is to retrieve the information that is more relevant to user query. The optimization technique used here is Genetic Algorithm (GA). A genetic algorithm is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and searchproblems. Genetic algorithms belong to the larger class of evolutionary algorithms, which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

In this work the Document Crawler is used for gathering and extracting information from the documents available from online databases and other databases.Since search space is too large, Genetic Algorithm (GA) is used to find out the combination terms. In the proposed document retrieval system, need to extract the keywords from the document crawler and with these keywords GA generate combination terms. To obtain better combination terms, need to calculate the fitness function based on the frequencies of keyword. The keyword frequency can be generated by counting occurrences of word in a document.

1. Introduction

Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining, a branch of computer science is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with Database Management. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices such as marketing, surveillance, fraud detection, and scientific discovery.

Genetic algorithms are unique in the sense that they do not optimize a problem by directly applying transformation operations on the physical representation of a problem. It applies transformations to its chromosome or genotype only.

Since GA works on a population of solutions it has the intrinsic ability to explore various areas of the search space concurrently. It exploits its current search information to intensively explore those regions of the search space that may quickly give good solutions and abandons other regions that are unlikely to yield good solutions. This adaptive feature of a GA enables it to be more intelligent than conventional algorithms.

As information has been increasing enormously in the world, it is difficult to retrieve the proper information as per the user satisfaction. In order to optimize the search result, the genetic algorithm (GA) is used. The document crawler is used for gathering and extracting information from the documents available from online databases and other databases. Since search space is too large, Genetic Algorithm (GA) is used to find out the best combination terms. Thus a fitness function is evaluated to initialize a population. After initializing the population, several modules of genetic algorithm is applied to get the optimized result. The population concept in genetic algorithm will result in variety of output for varying population. In order to check the efficiency of Genetic Algorithm various traditional methods are compared with the Genetic Algorithm.

Statement of Problem: In the existing system, they use hill climbing algorithm for document retrieval. In this methods of query expansion manipulate each term independent of other. In this methods of relevance feedback are not efficient when no relevant documents are retrieved with the initial query. This method used term co occurrences for query expansion in retrieving relevant documents where a term co occurrence deals with the terms that have related meaning. But here the difficulties like construction of a thesaurus and finding all the terms which has related meaning exists. Term co-occurrence, involves natural processing task like construction of a thesaurus by considering the semantics of the query, frequency characteristics of the terms and the related neighboring terms. Generally, the set of keywords are subdivided into classes of similar terms and treating the members of the same class alike for document retrieval. So use of co-occurrence data is having disadvantages.

2. Related work

Lena Tenenboim*et al*, [1] proposed a novel technique on ontology based classification. They have discussed on

classification of news items in ePaper, a prototype system of a future personalized newspaper service on a mobile reading device. By considering the difficulty that classical Euclidean distance metric cannot create an suitable separation for data lying in a manifold, a GA based clustering method with the help of geodesic distance measure is proposed by Gang Li *et al,* [2]. In the proposed method, a prototype-based genetic illustration is used, where every chromosome is a sequence of positive integer numbers that indicate the k-medoids. In addition, a geodesic distance based proximity measures is applied to find out the similarity between data points. Casillas*et al,* [3] put forth a novel concept on document clustering using GA. Andreas *et al,* [4] discussed on the clustering technique for text data. Text clustering usually involves clustering in a high dimensional space that appears complex with considered to virtually all practical settings. A Wordsets based document clustering algorithm forlarge datasets was proposed by Sharma *et al*., [5]. Cao *et al*., [6] provided fuzzy named entity-based document clustering. Conventional keyword-based document clustering methods have restrictions because of simple treatment of words and rigid partition of clusters. Zhang *et al*., [7] gives clustering aggregation based on GA for documents clustering. Web document clustering using document index graph is put forth by Momin*et al*., [8]. Muflikhah*et al*. [9] proposed a document clustering technique using concept space and cosine similarity measurement. Affinity-based similarity measure for Web document clustering is presented by Shyu*et al*., [10]. ELdesoky*et al*., [11] given a novel similarity measure for document clustering based on topic phrases. Thangamani et. al examined document clustering in individual and peer to peer environment and also developed the system for automatic extraction and classification of document using fuzzy ontology with genetic algorithm.

2. Motivation

Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories. From the Commercial Point of View, lots of data is being collected and warehoused. From the Scientific Point of View, data collected and stored at enormous speeds (GB/hour) like microarrays generating gene expression data.

3. Methodology

The aim of this proposed work is to retrieve the relevant documents by using the best combination of the term list, given a set of document collections. The keywords that are extracted from the document crawler are stored for generating the combination terms. After obtaining the best combination of terms, it is applied to the information retrieval system to obtain more relevant documents. Genetic Algorithm identifies the combinations of the terms that optimize the objective function.

The main objective of the project is to effectively retrieve the document from the document collection based on the frequency of the words in the document.The optimization technique used here is genetic algorithm to generate combinations of words.

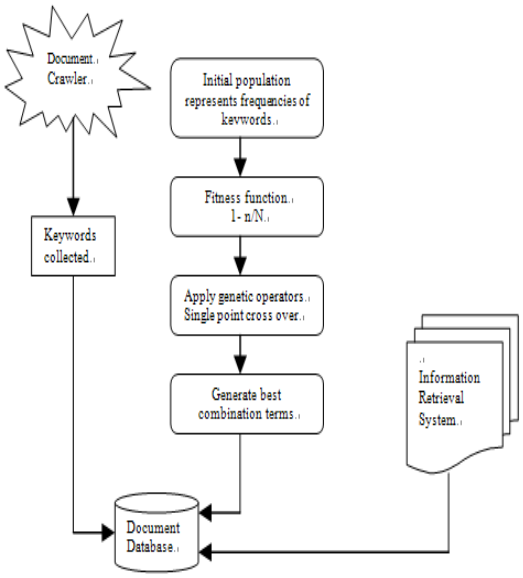

Figure: 3.1 System Architecture

This system is used to retrieve the relevant documents by using the best combination of the term list, given a set of document collections. The keywords that are extracted from the document crawler are stored for generating the combination terms. After obtaining the best combination of terms, it is applied to the retrieval system to obtain more relevant documents. Genetic Algorithm identifies the combinations of the terms that optimize the objective function. Based on the fitness function, generate the best combination terms. Then apply one point crossover for generating more combination of terms which need to optimize the result.

4. Evaluation Methods

The Genetic Algorithm parameters are initial population, fitness function, crossover and Mutation.

Initialize Population

Selection determines, which individuals are chosen for recombination (crossover).Initially many individual solutions are randomly generated to form an initial population. The population size depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions. Traditionally, the population is generated randomly; covering the entire range

of possible solutions (the search space).During each successive generation, a proportion of the existing population is selected to breed a new generation. This work uses, real time dataset collected from database which consists of keywords extracted by document crawler. The attributes available are id, keywords, frequency. The initial population is represented by randomly picking the term from the high frequency and the low frequency terms.
If any two terms are of same Frequency, then with that frequency, we add 0.01 to the frequency to maintain the uniqueness of the value for each keyword.

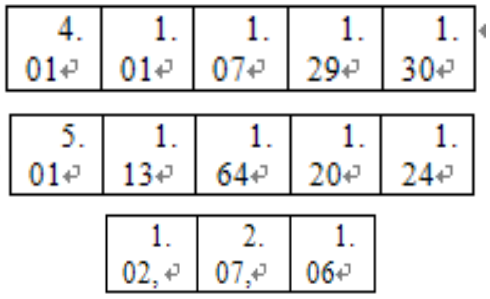

Figure:4.1 Different Chromosome representation

Crossover
Crossover is a genetic operator that combines two parents to produce offsprings. 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. For each new solution to be produced, a pair of "parent" solutions is selected for breeding from the pool selected previously. By producing a "child" solution using the methods of crossover, a new solution is created which typically shares many of the characteristics of its "parents". New parents are selected for each new child, and the process continues until a new population of solutions of appropriate size is generated. Crossover can be a fairly straightforward procedure. In this example,which uses the simplest case of crossover, two parents are randomly chosen to crossover. The cross over used is single point cross over.
The purpose of mutation is to preserve and introduce diversity. The newly created off springs may have the similar characteristics. In order to produce diversity among these off springsmutation is done by either changing the bits or multiplying it with constant values. This process will give a result of chromosomes that is different from the original.

Fitness Function
A fitness function is a particular type of objective function that prescribes the optimality of a solution (that is, a chromosome) in a genetic algorithm so that that particular chromosome may be ranked against all the other chromosomes. Optimal chromosomes, or at least chromosomes which are *more* optimal, are allowed to breed and mix their datasets by any of several techniques, producing a new generation that will be even better.
The fitness function used is,
Fitness function, F=1-n/N (1)
Wheren is the number of times the keywords are appearing in the whole document and N is the total number of documents present in the document collectionBased on the fitness function the combination terms are formed.

Termination Condition
This generational process is repeated until a termination condition has been reached. Common terminating conditions are: A solution is found that satisfies minimum criteria, fixed number of

generations reached and the highest ranking solution's fitness is reaching or has reached a plateau such that successive iterations no longer produce better.

5. Experiment Result
The existing system and proposed system are compared based on the efficiency in document retrieval, database size it occupies etc.

Characteristics	**Our Search Engine**	**Search Engines without Genetic Algorithms**
Document Relevancy	High	Less
Data Base Size	Very Less	Very High
Time Consumption	Variable	Constant
Search Space Coverage	More	Less

Table: 5.1 Comparison With Other Search Engines

6.Conclusion
The Traditional method gives less accuracy, whereas genetic algorithm provides more accuracy. Thus, there is a high rate

of optimization found in genetic algorithm with respect to change in population size. The various experiments conducted by the real time data set collected from the database were compared with the optimized results. Thus, it is clearly proven from the experimental results that Genetic

algorithm gives high rate of accuracy compared with the traditional methods. Finally concluding, it is obviously proven that genetic algorithm is highly efficient in optimizing document retrieval system.

This project will be implemented as a part of search engine for document retrievals only in the near future. We are aimed to fulfill the lagging part of this project to decrease the search time and make the different User Interfaces to the End User.

References

i. *Lena Tenenboim.,BrachaShapira. andPeretzShoval., "Ontology-Based Classification of News in an Electronic Newspaper", International Book Series Information Science and Computing, Pp: 89-98, 2008.*

ii. *Gang Li., JianZhuang., HongningHou. andDehong Yu., "A Genetic Algorithm based Clustering using Geodesic Distance Measure", IEEE International Conference on Intelligent Computing and Intelligent Systems, Pp: 274 - 278, 2009.*

iii. *Casillas, A., Gonzalez de Lena, M.T. and Martnez, R., "Document Clustering into an Unknown Number of Clusters Using a Genetic Algorithm", Lecture Notes in Computer Science, Vol.* ***2807****, Pp. 43-49, 2003.*

iv. *Andreas Hotho., Alexander Maedche. and Steffen Staab.,"Ontology-based Text Document Clustering", Journal on KunstlicheIntelligenz, Vol.* ***4****, Pp. 48-54, 2002.*

v. *Sharma, A. and Dhir, R., "A Wordsets based Document Clustering Algorithm for Large datasets", Proceeding of International Conference on Methods and Models in Computer Science, 2009.*

vi. *Cao, T.H., Do, H.T., Hong, D.T. and Quan, T.T.,"Fuzzy Named Entity-Based Document Clustering", IEEE International Conference on Fuzzy Systems, Pp. 2028 - 2034, 2008.*

vii. *Zhenya Zhang., Hongmei Cheng., Shuguang Zhang., Wanli Chen. andQiansheng Fang., "Clustering Aggregation based on Genetic Algorithm for Documents Clustering", IEEE Congress on Evolutionary Computation, Pp. 3156 - 3161, 2008.*

viii. *Momin, B.F., Kulkarni, P.J. and Chaudhari, A., "Web Document Clustering Using Document Index Graph", International Conference on Advanced Computing and Communications, Pp. 32 - 37, 2006.*

ix. *Muflikhah, L. and Baharudin, B., "Document Clustering Using Concept Space and Cosine Similarity Measurement", International Conference on Computer Technology and Development, Vol.* ***1****, Pp. 58-62, 2009.*

x. *Shyu, M.L., Chen, S.C., Chen, M. and Rubin, S.H., "Affinity-based similarity measure for Web document clustering", IEEE International Conference on Information Reuse and Integration, Pp. 247 - 252, 2004.*

xi. *Thangamani .M and Thangaraj P,"Ontology Based Fuzzy Document Clustering Scheme", Modern Applied Science,vol.4(7),2010,pp.148-153.*

xii. *Thangamani .M and Thangaraj .P,"Survey on Text Document Clustering",International Journal of Computer Science and Information Security,vol.8(4),2010.*

xiii. *Thangamani, M. and Thangaraj, P. "Integrated Clustering and Feature Selection Scheme for Text Documents", International Journal of Computer Science, Vol.6, Issue 5, pp.536-541, 2010.*

xiv. *Thangamani.M and Thangaraj.P, "Effective fuzzy semantic clustering scheme for decentralized network through multidomain ontology model", International Journal of Metadata, Semantics and Ontologies, Interscience Vol.7, Issue 2, pp.131-139, December 2012 Interscience publication*

xv. *Thangamani.M and Thangarj.P. "Fuzzy ontology for document clustering based on genetic Algorithm", International Journal of Applied mathematics and information science, Vol.4, Issue 7, pp.1563-1574, 2013.*