International Journal of Research in Advent Technology, Vol.2, No.2, February 2014

E-ISSN: 2321-9637

Optimization Techniques for Improving the Performance of Information Retrieval System

Yogesh Bade, Rutuja Bhat, Poonam Borate
Computer Department, Pimpri Chinchwad College of Engineering,Nigdi 44
byogesh26.4@gmail.com

Abstract— This article presents the new approch towards the Information Retrieval system for improving its performance and also states the technique for optimization. In Information retrieval System, effectiveness is achieved using genetic algorithm based on matching function . Genetic Algorithms (GA) is robust and efficient search technique and it can be used to search a user solution from large set of text document. Information retrieval is based on the similarity measurement between query and document. Text document with high similarity to query are match first and which are more relevant to the query should be retrieved first. In genetic algorithms, each query is represented by a chromosome. Also precision and recall is used to get improved result. In this paper we present the review of research in the field of Information Retrieval.

Keywords— Clustering, Informtion Retrieval(IR),Genetic algorithm(GA),Matching function, Overall matching function(OMF)

I. INTRODUCTION

Information Retrieval System (IRS) is used to store set of information that need to be processed and generate a ranking which reflect relevance between query and retrieves information corresponding to user's query. The goal of an Information retrieval system is to help the user to locate the relevant documents that have potential to satisfy the user's query. An Information Retrieval System consists of a software program that helps the user to find information as per their needs. IRS have to extract the keywords from the text documents and assign weights for each keyword. A matching techniques are used to match query with set of document. Thus, set of documents, the query and matching functions are important components of IRS. The information retrieval efficiency measures from recall and precision.

The proposed work aims at applying matching function. For this we have implemented existing matching functions and our work will be based on using GA for these matching functions.

Basic components of GA are:

A. Chromosome Representation

Chromosomes are the initial input given to GA. All the documents and query are first converted into chromosome. This is given as input to the genetic algorithm.

B. Fitness Function

Fitness Function gives a value which is used to calculate the similarity between query and document. Based on this value chromosome is selected for selection mechanism.

C. Selection operator

Selection is the process in which chromosomes is selected for next step or generation in genetic algorithm based on fitness value of chromosomes. Poor chromosome or lowest fitness chromosome selected few or not at all.

D.Crossover operator

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.

E. Mutation operator: Mutation is a process in which gene of the chromosome is changed. In one point mutation if gene is 0 then change it into 1 and if gene is 1 then change it into 0.

II. Related work

In this section we will briefly review research related to our work in this paper. Also we will survey the different research paper related to Information Retrieval system.

The Genetic-based approaches in ranking function discovery and optimization in information retrieval — A framework was proposed by author[1].
In this paper the main focus is given on ranking function and with the help of them how the optimization process is done.Firstly they focus on Vector Space Model which is the theoretical model upon which proposed integrated framework of ranking function discovery and optimization is based.
It is based on vector space and hence is easily interpreted from a geometric perspective. Each document and query is placed in an n dimensional space where its properties can be studied using geometrical similarity. This model has been one of the most successful models in various performance evaluation studies and most existing search engines and information retrieval systems are designed based on it.

In VSM, both documents and queries are represented as vectors of terms. Then the retrieval status value (RSV), is calculated for each document in the collection and the documents are ordered and presented to the user in the decreasing order of RSV. Various content based features are available in VSM to compute the term weights. The most common ones are the term frequency (tf) and the inverse document frequency (idf). Term frequency measures the number of times a term appears in the document or the query. The higher this number, the more important the term is assumed to be in describing the document. Inverse document frequency is calculated as log(N / DF), where N is the total number of documents in the collection, and DF is the number of documents in which the term appears. A high value of idf means the term appears in relatively few number of documents and hence the term is assumed to be important in describing the document.

GA[2] and GP[3] are artificial intelligence search algorithms based on evolutionary theory. They represent the solution to a problem as a chromosome (or an individual) in a population pool. They evolve the population of chromosomes in successive generations by following the genetic transformation operations such as reproduction, crossover, and mutation to discover chromosomes with better fitness values. A fitness function assigns the fitness value for each chromo-some and represents how good the chromosome is at solving the problem at hand.

Because of the intrinsic parallel search mechanism and powerful global exploration capability in a high-dimensional space, both GA and GP have been used to solve a wide range of hard optimization problems. GAs are typically used to solve difficult parameterized nonlinear optimization problems, while GP is typically used to approximate or discover complex, nonlinear functional relationships.

The Minimizing time risk in on-line bidding : An information retrieval based approach was proposed by author[4].
In this paper the main focus is given on the e-bidding applications. As Knowledge is of prime importance, particularly for the individuals who are involved in e-business. A lot of Energy and time is wasted by the individuals in seeking required knowledge and information. In order to Facilitate the individuals with required information, an efficient technique for the proper retrieval of Knowledge is must. Almost all online business activities, particularly e-auction based firms are surrounded by various risk factors pertaining to time, security, brand etc. The main focus of the present paper is to analyze all such risk factors and further to categorize the same as per their degree of influence.

The contribution in the field of information retrieval domain is mainly categorized on the basis of solution methodologies employed to achieve certain objectives. Solution methodologies that are prevalent is information retrieval are:
_ Probabilistic information retrieval (Fuhr & Buckley, 1991; Gordon,
1988).
_ Knowledge based information retrieval (Chen & Dhar, 1991).
_ Learning based information retrieval (Chen, 1995; Yang & Korfhage,
1993).

Auction is an economic incentive mechanism that determines the price of an item to buy or sell (Huang, Tseng, & Kusiak, 2002). This process requires the involvement of one or more bid-ders who want the item, and an item for sale to the highest bidder. Online auction is done through internet, where, a web page used to display information regarding goods or services along with the information to sell them.
Traditional auctions have some information deficiencies, as they last only for a few minutes for each item sold. In this auction type, both sellers and bidders may not get what they actually want. With the advancement in the information technology, auction can be performed via internet, thus, overcoming the difficulties related to available information.

The various Risk factors such as Time risk, Security risk, Vendor's risk, Price collusion, Brand risk, Privacy risk etc is analyzed and then processed for a nominal group technique (NGT) based approach has been utilized to resolve the underlying task (refer Lai, Ho, & Chang, 1998). A committee of five members are formed to implement the same in present re-search. We have modified the NGT technique by categorizing it into three broad phases termed viz. selection phase, screening phase, and implementation phase. Selection phase involves the collection of pertinent information by the committee members to further generate a list of items. In present context, this process is equivalent to the collection of information pertaining to risk fac-tors in online bidding process and further, based upon the acquired information, the concerned list generated by them is recorded/ stored separately. Once the collected set of items has been properly recorded, the screening phase starts that aims to eliminate the poor conceptualized items and to promote stronger points. After this, a final list of the risk factors is listed, which acts as input for application phase. Among the final recorded list of risk factors one of them is selected to determine its importance level.

A Genetic Algorithm is an 'intelligent' probabilistic search algorithm that simulates the process of evolution by taking a population of solutions and applying genetic operators in each reproduction. Each solution in the population is evaluated according to some fitness measure. Fitter solutions in the population are used for reproduction. New offspring' solutions are generated and unfit solutions in the population are replaced. The cycle of evaluation–
selection–reproduction is continued until a satisfactory solution is found Goldberg (1989) and Michalewicz (1994). Holland(1975) introduced genetic algorithms which, later on, were applied to a wide variety of problems.

The survey paper on various methods in content based information retrieval was proposed by author[5].
In this paper the focus is given on web baed IR system.Information Retrieval is an emerging research area in the field of Information Retrieval. Due to the immense
amount of data in the WWW, it is very tough for the user to retrieve the relevant images. Traditional Image Retrieval
approaches based on topic similarity alone is not sufficient nowadays the content based image retrieval (CBIR) are
becoming a source of exact and fast retrieval. A variety of techniques have been developed to improve the performance of CBIR. Data clustering is an unsupervised method for extraction hidden pattern from huge data sets. With large data sets, there is possibility of high dimensionality. Having both accuracy and efficiency for high dimensional data sets with enormous number of samples is a challenging arena. In this paper the clustering techniques are discussed and analysed.

Clustering techniques can be classified into supervised (including semi-supervised) and unsupervised schemes. The former consists of hierarchical approaches that demand human interaction to generate splitting criteria for clustering. In unsupervised classification, called clustering or exploratory data analysis, no labeled data are available The goal of clustering is to separate a finite unlabeled data set into a finite and discrete set of "natural," hidden data structures, rather than provide an accurate characterization of unobserved samples generated from the same probability
distribution This paper critically reviews and summarizes different clustering techniques.

Images can be clustered based on the retrieval system logs maintained by an information retrieval process. The session keys are created and accessed for retrieval. Through this the session clusters are created. Each session cluster generates log –based document and similarity of image couple is retrieved. Log –based vector is created for each session vector based on the log-based document. Now, the session cluster is replaced with this vector. The unaccessed documents creates its own vector. Hierarchical clustering (HC) algorithms organize data into a hierarchical structure according to the proximity matrix. The results of HC are usually depicted by a binary tree or dendrogram where A, B, C, D, E, F, G are objects or clusters. It represents the nested grouping of patterns and similarity levels at which groupings change. The root node of the dendrogram represents the whole data set and each leaf node is regarded as a data object. The intermediate nodes, thus, describe the extent that the objects are proximal to each other; and the height of the dendrogram usually expresses the distance between each pair of objects or clusters, or an object and a cluster.

A rough classification retrieval system is formed. Retrieval Dictionary Based Clustering is formed by calculating the distance between two learned patterns and these learned patterns are classified into different clusters followed by a retrieval stage. The main
drawback addressed in this system is the determination of the distance. To overcome this problem a retrieval system is
developed by retrieval dictionary based clustering. This method has a retrieval dictionary generation unit that classifies

learned patterns into plural clusters and creates a retrieval dictionary using the clusters.

K Means Clustering is nonhierairchal method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart. Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated everytime a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.The K- means algorithm is very simple and can be easily implemented in solving many practical problems. It can work very well for compact and hyperspherical clusters. The time complexity of K-means is O (NKd).

III.PROPOSED IR SYSTEM USING GENETIC ALGORITHM

Genetic algorithms are adaptive heuristic search algorithm. Genetic algorithm is based on evolutionary ideas of natural selection and genetics. Genetic algorithm is often used to solve problems and looking for best solution. In genetic algorithm query which is entered by user and text documents which are stored in data repository represented as chromosome. Each chromosome has specific fitness value. Population is generated from the set of chromosome and their associated fitness. There are some genetic operator such as selection ,crossover , mutation. These genetic operator are used to generate new population from existing population. It's a genetic algorithm based model for information retrieval. In this all the matching functions are act as fitness function for GA.

Algorithm

1. User enters query into our system.
2. Match keywords from user query with list of keywords
3. Preprocess keywords and text documents.
4. Encode documents retrieved by user query to chromosomes (initial population)
5. Population feed into genetic operator process such as selection, crossover, and mutation.
6. Repeat step 5 until maximum generation is reached. We will get an optimize query chromosome for text document retrieval.
7. Decode optimize query chromosome to query and retrieve document from database.

GA model for IR

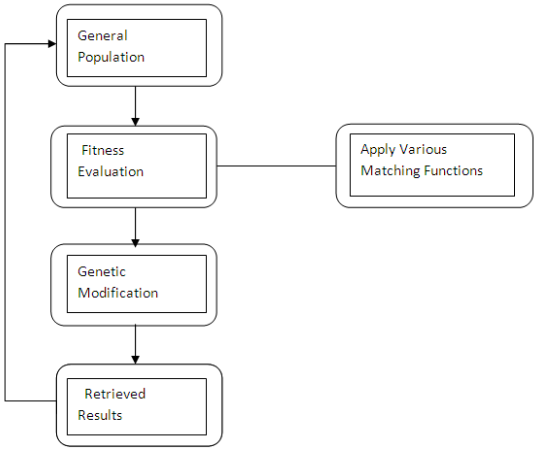

Fig.: GA model for Information Retrieval

Performance Measures:

Recall is defined as the proportion of relevant document retrieved.

It is the fraction of the document s that are relevant to the query that is successfully retrieved.

recall=(|{relevant documents}∩{retrieved documents}|) / (|{relevant documents}|)

Precision is defined as the proportion of retrieved document that is relevant.

It is the fraction of the documents retrieved that are relevant to the user's information need.

Precision=(|{relevant documents}∩{retrieved documents}|) / (|{retrieved documents}|)

The Basic components of GA are:

- Chromosome Representation
- Fitness Function
- Selection operator
- Mutation operator
- Crossover operator

IV MATCHING FUNCTIONS

1. **Classical matching function**

Fitness function is used to measure the performance of solution. It evaluate how solution is good. We also use the fitness functions to calculate the distance between document and query.

We define X = (x1 , x2 , x3 ,….., xn) , | X | = number of terms occur in X

Result from these fitness functions are interval 0 to 1. By 1.0 means document and query is sameness. Values near 1.0 mean documents and query are more relevant and values near

0.0 mean documents and query are less relevant. Values evaluate from fitness functions are called “fitness”.

We are using following matching functions:

1.Dice Coefficient:

$$\frac{2\sum_{i=1}^{t} xi * yi}{\sum_{i=1}^{t} xi^2 + \sum_{i=1}^{t} yi^2}$$

2.Jaccard Coefficient:

$$\frac{\sum_{i=1}^{t} x\,i * yi}{\sum_{i=1}^{t} xi^2 + \sum_{i=1}^{t} yi^2 - \sum_{i=1}^{t} xi * yi}$$

3.Cosine Coefficient:

$$\frac{\sum_{i=1}^{t} x\,iyi}{\sqrt{\sum_{i=1}^{t} xi^2 * yi^2}}$$

V CONCLUSION

The proposed information retrieval system is more efficient within a specific domain as it retrieves more relevant results.. A genetic algorithms offer domain independent search ability. The capabilities of GA are used in the proposed work to improve the performance of information retrieval.

REFERENCES

[1] Genetic-based approaches in ranking function discovery and optimization in information retrieval — A framework-Weiguo Fan [a], Praveen Pathak [b], , Mi Zhou [c]

[2]J.H. Holland, Adaptation in Natural and Artificial Systems, 2nd ed.MIT Press, 1992.

[3] J.R. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, MA, USA, 1992.

[4] Minimizing time risk in on-line bidding: An adaptive information retrieval based approach-Anoop Verma [a], M.K. Tiwari [b], Nishikant Mishra [c]

[5] Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, and Osman A. Sadek,” Using Genetic Algorithm to Improve Information Retrieval Systems”, 2008.

[6] A.Thakare, Dr.C.A. Dhote, ”Virtual center based Algorithms for information retrieval”, interanational Journal of Engineering research & Applications, volume 2, Issue-6,Nov-Dec,2012 Page no-1312-1316

[7]A.Thakare, Dr.A.C. Dhote ,” Evolutionary Approach for effective Clustering and IR”, ICCTD-2011,Chengdu,china.

[8] A. S.Siva Sathya, “A Document Retrieval System with Combination Terms Using Genetic Algorithm “, 2010.

[9] Mr. Vikas Thada, Mr. Sandeep Joshi.”A genetic algorithm approach for improving the average relevancy of retrieved documents using Jaccard Similarity Coefficient”, August, 2011.

[10]Vector Space Model,” Vector Space Model”, 2003.

[11] Huda Yasin, Mohsin Mohammad Yasin, Farah Mohammad Yasin, “Automated Multiple Related Documents Summarization via Jaccard’s Coefficient”, January 2011.