Genetic Algorithm for Information Retrieval

Philomina Simon
Ramanujan School of Mathematics and Computer Science,
Department of Computer Science ,
Pondicherry University,
Puducherry, India
philominasimon@yahoo.com

S. Siva Sathya
Ramanujan School of Mathematics and Computer Science,
Department of Computer Science ,
Pondicherry University,
Puducherry, India
sivasathya@yahoo.com

Abstract— **Retrieval of relevant documents from a collection is a tedious task. As Genetic Algorithms (GA) are robust and efficient search and optimization techniques, they can be used to search the huge document search space. In this paper, a general frame work of information retrieval system is discussed. The applicability of Genetic algorithms in the field of information retrieval is also discussed. A review on how GA is applied to different problem domains in information retrieval is presented. A study on the similarity measures in information retrieval is also discussed.**

Keywords-Genetic Algorithm; Information Retrieval; Query Optimization/; Query Formulation; Genetic Mining; Document Indexing; Web Search; Ranking; Match Function Adaptation

I. INTRODUCTION

The goal of an Information Retrieval System (IRS) is to help a user to locate the most similar documents that have the potential to satisfy the user information needs. The focus of information retrieval is the ability to search for information relevant to a user's needs within a collection of data which is relevant to the user's query. User formulates a query and sends to the information retrieval system. Information retrieval system searches for the matches in the document database and retrieves the results. The user evaluates the results based on the relevance. This survey includes different proposals found for the application of genetic algorithm to the field of information retrieval. Different kinds of IR problems that are solved by genetic algorithms are analyzed.

Section 2 reviews the applicability of genetic algorithm in different fields of information retrieval. Section 3 covers the similarity measures dealt in information retrieval. Section 4 gives the conclusion.

A. An Information Retrieval System Framework

The three main components of an information retrieval system is shown in fig-1 [21]. It is composed of Documentary Database, Query Subsystem and Matching Mechanism. Documentary database stores the documents and their representations. This component also contains an indexer module which automatically generates a representation for each document by extracting the document contents.

Query Subsystem does query formulation. This component allows the user to formulate the queries. It contains a query language that collects the rules to generate queries. This component contains some functions to select the relevant documents.

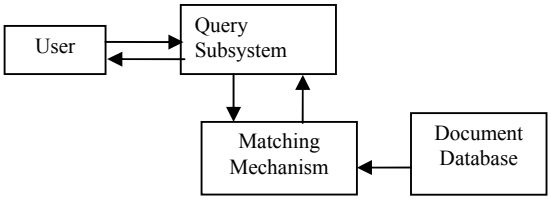

Fig: 1 Information Retrieval System framework

Matching Mechanism compares the set of documents in the document database with the query which is given by the user. The documents which match with the query given are termed as relevant documents. So this component helps to retrieve the relevant documents. It analyzes the performance to extend which document representations satisfy the query and retrieve the documents evaluated as relevant.

B. Usage of Genetic Algorithm in Information Retrieval

There are three main components that have to be taken care while designing GA [1]. The first one is coding the problem solutions, next is to find a fitness function that can optimize the performance and finally, the set of parameters including the population size, genetic operators and their percentages. Genetic algorithms are a powerful search mechanism and it is suitable for the information retrieval for the following reasons [16].The document search space represents a high dimensional space. GAs are one of the powerful searching mechanism known for its robustness and quick search capabilities. So they are suitable for information retrieval. In comparison with the classical information retrieval models, GA manipulates a population of queries rather than a single query. Each query may retrieve a subset of relevant documents that can be merged. The terms occur in documents as groups. GA contributes to maintain useful information links representing a set of terms indexing the relevant documents.

978-1-4244-4711-4/09/$25.00 ©2009 IEEE

C. Genetic Model For Information Retrieval

Information Retrieval (IR) may be defined, in general, as the problem of selection of document information from storage in response to search questions provided by a user. Information retrieval system (IRS) deals with document databases containing textual, pictorial or vocal information and process user queries trying to allow the user to access relevant information in an appropriate time interval.

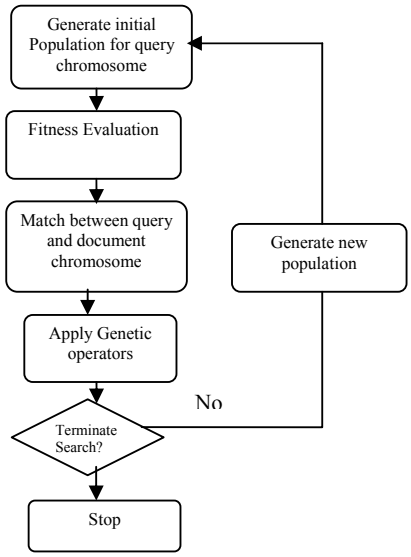

Fig.2. Steps of Genetic Algorithm

This information retrieval approach given in Fig 2 can be applied for retrieving information. Query and documents are represented as chromosome. An initial population of query is created. The query is sent to the information retrieval system, a match is done between the query chromosome and the document chromosome, and then the document is considered as relevant. If non relevant documents are found, then the query is reformulated. The query is reformulated until a relevant document is retrieved. The document collection consists of many documents containing information about various subject or topics of interests. The primary concern in representation is how to select proper index terms. Representation proceeds by extracting the key words that are considered as content identifiers and organizing them into the required format.

II. Applicability of Genetic Algorithms in Different Areas of Information Retrieval

Research work is going on in the field of information retrieval to make it effective. The main issues to be discussed are obtain more pages relevant to the user's query , optimize the search time. GA is a stochastic search algorithm which tries to optimize the solution. In the field of information retrieval, GA has been used widely to optimize the query and obtain a relevant set of pages. GAs have been applied to solve some of the information retrieval problems. The problem areas include Genetic Mining, Agents for internet search, Query Formulation, Query Optimization, Document Indexing, Ranking, Document Clustering, Match function Adaptation, Rough Sets.

A. Genetic Mining

Genetic mining is the mining of data by genetic algorithms to improve the efficiency of the document retrieval. Web documents have a number of tags indicating the structure of texts. *Sun Kim [2]* proposed a genetic algorithm that learns the internal structure of HTML documents which are used to re rank the documents retrieved by standard weighing schemes for improving the performance. The use of document structure mining approach tends to move relevant documents to upper ranks, which is necessary in web information retrieval systems. *Ardil* [20] proposed another genetic algorithm in concept weighting and topic identification, based on the concept of standard deviation The term "discourse topic" is used as a representation of the document content in order to distinguish it from other documents in corpus. Discourse topic is an understandable topic which allows a person to quickly see what the topic is about. Concept is used to state some related words that point to a specific entity or impression in a document.

B. Agents for Internet Search

Different proposals are put forward by the authors that use GAs in Internet search to make the search process easier. In [3], an intelligent personal spider approach for Internet searching is proposed. Chen et al. implemented Internet personal spiders based on best first search and GA techniques. The used GA applies stochastic selection based on Jaccard's fitness, with heuristic-based crossover and mutation operators. These personal spiders dynamically take a set of user's selected starting homepages and search for the most closely related homepages in the web, based on the existing links and keyword indexing.

C. Query Optimization

Query optimization is defined as the procedure or procedures used to make an information retrieval system effective by improving the query using mathematical techniques.

A query with Boolean and logical operators was used in information retrieval [22]. For genetic algorithms, encoding chromosomes was done from Boolean query, where it was represented in the form of tree prefix with indexing for all terms and all Boolean logical operators. Information retrieval effectiveness measures precision and recall were used as a fitness function in the work. Other Genetic algorithms operators used were as single point crossover on Boolean logical operators, and mutation operator was used to exchange one of the Boolean operators AND, OR and XOR with any other one.

Rocı´o L. Cecchini , Carlos M. Lorenzetti , Ana G. Maguitman, Ne´lida Beatrı´z Brignole proposed optimization techniques based on Genetic Algorithms to evolve ‘‘good query terms” in the context of a given topic [19].These techniques place emphasis on searching for novel material that is related to the search context. The use of a mutation pool to allow the generation of queries with new terms, study the effect of different mutation rates on the exploration of query-space is discussed and the use of a specially developed fitness function that favors the construction of queries containing novel but related terms. This proposal is a novel GA approach to Web search based on thematic contexts. Proposed method is fully automatic and does not require relevance feedback from the users.

D. Document Indexing

Indexing is the process which transforms an unstructured set of tokens, such as words or terms, to a data structure, called index. An index is a representation of the content of a collection of documents and of each document. It is by means of indexing that tokens that are found in documents individually or in groups become key words, index terms, or descriptors, thereby assuming the representational power that is needed in order to identify relevant documents. Given an input query, an IR system accesses the indexes and selects the documents probably relevant to the end user’s information need represented by that query. In the field of document indexing, the early works were done by Gordon. In his work, he proposed to associate more than a single description to each document and to adapt them throughout time as a good solution to the problem of the different forms that different users’ queries searching for the same documents can present.

Gordon [21, 18] proposed a GA to derive the document descriptions. He chose a binary coding scheme where each description is a fixed length, binary vector. The genetic population is composed of different descriptions for the same document. The fitness function is based on calculating the similarity between the current document description and each of the queries by means of the Jaccard’s index, and then computing the average adaptation values of the description to the set of relevant and non-relevant queries.

Vrajitoru worked with the vector space model, and thus document descriptions are real vectors in the [0, 1] interval which represent the weights associated to each term [17]. Each document has associated just one description which leads to encode the whole collection in a single chromosome. The main problem of this model is that the fitness function only considers one query, and thus the document descriptions are adapted to match with this single query and not with a set of queries as in Gordon’s model.

E. Document Clustering

While query expansion aims at expanding the number of relevant documents attracted by a given query, there are some other retrieval techniques that discover relationships among hopefully relevant documents independently of any query. An example of these techniques is document clustering. These techniques base their effectiveness on the Cluster Hypothesis: the documents which are relevant to the same query tend to be strongly related to each other. Robertson and Willet’s [11] proposed the idea to look for groups of terms appearing with similar frequencies in the documents of a collection

In [13], Gordon designs a philosophy according to which it is possible to make a user-oriented clustering of documents using any classical clustering technique. The basic idea is that the system adapts document descriptions throughout time. In this way, documents being relevant to a query finally have similar descriptions and the system periodically clusters the new adapted descriptions. Gordon worked with the data base used in and considers the relative density of the cluster as goodness measure.

F. Query Formulation

Query formulation is an essential part of successful information retrieval. Query formulation deals with how well we are framing the user query. The three main factors affecting query formulation are media expertise, domain expertise and type of search. One retrieval technique prominently used to improve recall is called query expansion. Query expansion takes a query as input and adds new terms to it thereby producing a new, expanded query. The expansion terms may originate directly from thesauri, dictionaries, or end users.

Abdelmgeid A.Aly [4] proposed adaptive method using genetic algorithm to modify user’s queries, based on relevance judgments. The algorithm shows the effects of applying GA to improve the effectiveness of queries in IR systems. The goal is to retrieve most relevant documents with less number of non-relevant documents with respect to user's query in information retrieval system using genetic algorithm. The proposed GA approach gives better results than classical IR system when tested.

Chen [8] et al. used a GA as an Inductive Query By Example (IQBE) technique to learn the query terms ie for team learning that better represent a relevant document set provided by a user. They consider an information retrieval system based on the vector space model. Each chromosome is a fixed size, binary vector where each position is associated to an existing term in the initial relevant document set. Genetic operators are one-point crossover, uniform mutation and roulette wheel selection.

Hao Lang, Bin Wang, Gareth Jones, Jin-Tao Li, Fan Ding and Yi-Xuan[10] Liu proposed a new statistical method called Covering Topic Score (CTS) to predict the performance of the query for information retrieval. Estimation is based on how well the topic of user’s query is covered by documents retrieved from a certain retrieval system.

The goal of Aris Anagnostopoulos , Andrei Broder , Kunal Punera [15] is to build the “best” short query that characterizes a document class using operators available within search engines. Good classification accuracy can be achieved on average over multiple classes by queries with as few as ten terms. The proposed work shows that optimizing the efficiency of query execution by careful selection of terms can

further reduce the query costs and also how classification can be performed effectively and efficiently using a search-engine model.

G. Match Function Adaptation

The aim is to use an GA to generate a similarity measure for a vector space information retrieval system to improve its retrieval efficacy for an specific user. This constitutes a new relevance feedback since matching functions are adapted instead of queries. In [6], Pathak et al. propose a new weighted matching function, which is the linear combination of different existing similarity functions. The weighting parameters are estimated by a GA based on relevance feedback from users. They used real coding, a classical generational scheme, two-point crossover and Gaussian noise mutation.

H. Rough Sets

Rough Genetic Algorithm [7] has its application in Information Retrieval. It is based on the concept of rough values. The rough value consists of an upper bound and a lower bound. Variables such as daily temperature are associated with a set of values instead of a single value. The upper and lower bounds of the set can represent variables using rough values. Rough equivalents of basic notions such as gene and chromosomes are defined. Two genetic operators in rough GA namely, *union* and *intersection* is also discussed which helps in the rough computing. These rough genetic operators provide additional flexibility for creating new generations during the evolution. Two new evaluation measures, called *precision* and *distance*, are also defined. The precision function quantifies information contained in a rough chromosome, while the distance function is used to calculate the dissimilarity between two rough chromosomes. A simple document retrieval example was used to demonstrate the usefulness of RGAs. Rough genetic algorithms seem to provide useful extensions for practical applications.

TABLE 1: EVALUATION OF GENETIC ALGORITHM IN DIFFERENT AREAS OF INFORMATION RETRIEVAL

Application Area	***Purpose of GA***	***Chromosomes***	***Fitness function***	***Genetic Operators***
Genetic Mining [21]	Learn the internal structure of HTML documents	Tag weights are encoded as chromosome	Fitness function f: J—>R measures the performance of retrieval results using tag weights.	Roulette selection Truncation selection One point cross over Uniform cross over Single Point Mutation
Agents for internet search [3]	Intelligent Searching	The search space of the problem is represented as a collection of individuals, which are referred as chromosomes.	Jaccards fitness function	Stochastic selection Heuristic based cross over and mutation operators
Query Optimization[75]	Retrieve more relevant documents with respect to user query	Chromosome encoding done with Boolean query where it is represented of a tree structure	Precision Recall	Single Point Cross Over Single Point Mutation
Match Function Adaptation[37]	Generate a new weighted matching function	Real coding	DCV(document cut off value) is calculated Precision Recall	Roulette Wheel selection Two point cross over Gaussian Mutation
Rough Sets[7]	Rough computing for finding a solution from a set of values	Rough Chromosomes and Rough genes	Precision Recall A distance measure for quantifying the did similarity between two rough chromosomes	Union Intersection

TABLE 2: COMPARISON OF DIFFERENT PROPOSALS THAT USE GENETIC ALGORITHM FOR INFORMATION RETRIEVAL

Different Proposals	***Purpose of GA***	***Chromosomes***	***Fitness function***	***Genetic Operators***
Document Indexing				
Gordon[21]	Derive document descriptors	Different descriptors for the same documents	Jaccard's Index Average adaptation values	Single point cross over No mutation operator
Vrajitoru[17]	Work with vector space model.	One document description which encodes whole collection in a single chromosome	Fitness function considers only one query Average Precision	Two point cross over Dissociated cross over
Document Clustering				
Robertson and Willet[11]	GA grouping the terms without maintaining the initial order	Two different coding schemes used. Separator method Division -assignment method	Measure of relative entropy Pratt's measure	Roulete Wheel Selection Order-based, position based, one point and two point cross over Inversion, random sublist and position mutation operators

Query Formulation				
Chen , SankaraNarayanan, She[8]	GA for IQBE technique to learn the query terms that better represent a relevant document	Fixed size, binary vector where position is associated to an existing term in the initial relevant document set.	Jaccards Score	Roulette wheel selection One point cross over Uniform mutation
Robertson and Willet[12]	Investigate upper bound for relevance feed back Steady State evolution model is used.	Two different coding schemes: Gray and Real. Initial Population is randomly generated	Measures the degree in which the weight vector maintains the optimal order of documents	One point cross over Two point cross over Random mutation operator Creep Mutation
Yang and Korfaghe	GA for weight learning	Real Coding	3 different fitness functions used based on Recall and Precision	Two point cross over Random mutation operator
Sanchez, Miyano[9]	Learn term weights of extended booolean queries for fuzzy IRS	Binary coded chromosomes encode n term weights and similarity threshold	Linear combination of Precision and Recall	One point cross over Single point Mutation
Horng and Yeh's [14,5]	Adapt to the query term weights in order to get the closest vector to optimal one Based on vector space model	Initial population randomly generated. Both query and documents are represented as vectors	Non interpolated average precision which takes care of order of appearance.	Weight selection and natural cross over Mutation operator is considered as inversion of a weight.

TABLE 3: COMPARISON OF SIMILARITY MEASURES IN INFORMATION RETRIEVAL

Fitness Function	***Score Calculation***	***Features***	***Application Areas***
Jaccard's Coefficient [21]	$\frac{\#(X \cap Y)}{\#(X \cup Y)}$	Finding similarity in documents. Used in Boolean Algebra Useful when negative values give no information.	Query optimization, Text Mining
Dice Coefficient[14]	$\frac{2\lvert X \cap Y\rvert}{\lvert X\rvert+\lvert Y\rvert}$	A correlation coefficient for discrete events. Dice is similar to Jaccard but gives twice the weight to agreements.	Text Comparison Conceptual Graph Matching Text Mining
Horng and Yeh Coefficient[14,5]	$F = \frac{1}{\lvert D\rvert}\sum_{i=1}^{\lvert D\rvert}\left(r(d_i)\sum_{j=1}^{\lvert D\rvert}\frac{1}{j}\right)$	The fitness with a higher score reflects a higher probability similarity of document.	Searching Retrieval of required document from one or more databases based on the similarity level

III. COMPARISON OF SIMILARITY MEASURES IN INFORMATION RETRIEVAL

The objective of GA was to find a set of documents which best fit the searcher's needs. The genetic algorithm performs better depending upon the fitness function applied to the problem. Here commonly used fitness functions in the domain of information retrieval like Jacccard's coefficient, Dice Coefficient and Horng and Yeh Coefficient are compared and . The comparison deals with the score calculation , features of the fitness function. The application areas in which these similarity measures are used are also discussed in this paper.

IV. CONCLUSION

This survey has dealt with the study the basics of the information retrieval and genetic algorithm The research areas in information retrieval and various issues that can be solved using the optimization and searching technique of GA. It also deals with the different application domains in information retrieval which are emerging research areas. This study discusses the applicability of genetic algorithm in different areas of information retrieval such as genetic mining, query optimization, document clustering, and query optimization etc.

REFERENCES

[1] Lothar M. Schmitt ," Fundamental Study ,Theory of genetic algorithms", Theoretical Computer Science 259 , 1–61, 2001.

[2] Sun Kim and Byoung, Genetic Mining of HTML structures for effective information retrieval , Applied Intelligence , 18, 243-256 , 2003

[3] H. Chen, C. Yi-Ming, M. Ramsey, C. Yang, "An intelligent personal spider (agent) for dynamic Internet/Intranet searching", Decision Support Systems 23 (1998) 41–58.

[4] Abdelmgeid A.Aly, Applying genetic algorithm in query improvement problem, International journal "Information Technologies and Knowledge" Vol 1, 309 – 316, 2007

[5] Horng, J.-T., Yeh, C.-C., Applying genetic algorithms to query optimization in document retrieval. Information Processing and Management 36(5), 737–759 ,2000

[6] P. Pathak, M. Gordon, W. Fan, Effective information retrieval using genetic algorithms based matching functions adaptation, in, Proc. 33rd Hawaii International Conference on Science (HICS), Hawaii, USA, 2000.

[7] Pawan Lingras ,"Unsupervised Rough Set Classification Using GAs ", Journal of Intelligent Information Systems Volume 16 , Issue 3 ,pp: 215 - 228 , August 2001

[8] H. Chen et al., A machine learning approach to inductive query by examples: an experiment using relevance feedback, ID3, genetic algorithms, and simulated annealing, Journal of the American Society for Information Science 49 (8) (1998) 693–705

[9] E. Sanchez, H. Miyano, J. Brachet, Optimization of fuzzy queries with genetic algorithms. Applications to a database of patents in biomedical engineering, in: Proc. VI IFSA Congress, Sao- Paulo, Brazil, 1995, pp. 293–296.

[10] Hao Lang, Bin Wang, Gareth Jones, Jin-Tao Li, Fan Ding and Yi-Xuan Liu , Query Performance Prediction for Information Retrieval Based on Covering Topic Score, Journal of computer science and technology 23(4),590- 601,2008

[11] A.M. Robertson, P. Willet, Generation of equifrequent groups of words using a genetic algorithm, Journal of Documentation 50 (3) , 213–232., 1994

[12] A. Robertson, P. Willet, An upper bound to the performance for ranked-output searching: optimal weighting of query terms using a genetic algorithm, Journal of Documentation 52 (4),1996 , 405–420.

[13] M. Gordon, User-based document clustering by redescribing subject description with a genetic algorithm, Journal of the American Society for Information Science 42 (5) 311–322., 1991

[14] Zhengyu Zhu, Xinghuan Chen, Qingsheng Zhu, Qihong Xie, A GA-based query optimization method for web information retrieval , Applied Mathematics and Computation 185 (2007) 919–930.

[15] Aris Anagnostopoulos · Andrei Broder · Kunal Punera ,Effective and efficient classification on a search-engine model , Knowl Inf Syst (2008) 16,129–154

[16] M. Boughanem, C. Chrisment, L. Tamine, "Multiple query evaluation based on an enhanced genetic algorithm", Information Processing and Management 39 , 215–231, 2003

[17] D. Vrajitoru, Crossover improvement for the genetic algorithm in information retrieval, Information Processing and Management 34 (4) ,405–415 ,1998

[18] O.Cordon, .E.Herrera , C. Lopez- Pujalte, M.Luque, C.Zarco ,A review on the application of evolutionary computation to information retrieval, International Journal of Approximate reasoning34, 241- 264, 2003.

[19] Rocı´o L. Cecchini , Carlos M. Lorenzetti , Ana G. Maguitman, Ne´lida Beatrı´z Brignole , Using genetic algorithms to evolve a population of topical queries , Information Processing and Management , 2008

[20] S. M. Khalessizadeh, R. Zaefarian, S.H. Nasseri, and E. Ardil , Genetic Mining, Using Genetic Algorithm for Topic based on Concept Distribution, PROCEEDINGS OF World academy of science , engineering and technology ,143 -147, 2006

[21] Michael Gordon, Applying probabilistic and genetic algorithms for document retrieval, Computer Practies, 1208 -1218 , 1988

[22] Suhail S.J Komer, Dsan Husek, Using genetic algorithms for Boolean optimization, Proceedings of Conference, IEEE,178-183