International Journal of Computer and Electrical Engineering, Vol. 2, No. 1, February, 2010
1793-8163

A Document Retrieval System with Combination Terms Using Genetic Algorithm

First A. S.Siva Sathya , Second B. Philomina Simon, *Member IACSIT*

***Abstract*—As information has been increasing enormously in the world, it is difficult to retrieve the proper information as per the user satisfaction. In our proposed work, 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, we extract the keywords from the document crawler and with these keywords GA generates combination terms. The proposed work is having three main features: First is to extract keywords and other information from the database by a document crawler. Second is to generate the combination terms using genetic algorithm. Third, results generated from the GA are applied to information retrieval system to generate better results. From the results obtained, the relevance of the documents are verified using evaluation measures namely precision and recall.**

***Keywords*—Combination terms, Document Retrieval, Genetic Algorithm, Document Crawler, keyword**

I. Introduction

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. The general objective of an Information Retrieval System is to minimize the overhead of a user locating needed information. Overhead can be expressed as the time a user spends in all of the steps leading to reading an item containing the needed information (e.g., query generation, query execution, scanning results of query to select items to read, reading non-relevant items).

An Information Retrieval System consists of a software program that facilitates a user in finding the information the user needs. IR is to provide the users with the documents that satisfy their information need. IRS have to extract the key words from the documents and assign weights for each keyword. The conditions that an IR system should follow to be effective are given as follows: The IRS should be scalable enough to be able to handle large document collections .It must be able to build the indexes in a reasonable amount of time to ensure the index efficiency. Query efficiency must be ensured to find out whether the queries are running fast. Query Effectiveness also affects the IRS since the retrieved result set must be relevant. Research in IR includes modeling, document classification and categorization, systems architecture, user interfaces, data visualization, filtering, languages, etc

The crawler which is a part of information retrieval system is used to gather the information from the websites; here it acts as a document crawler from which we derive the information content. An important aspect that is to be considered about crawlers is the nature of the crawl task. Crawl characteristics such as queries and/or keywords provided as input criteria to the crawler, user-profiles, and desired properties of the pages to be fetched can lead to significant differences in crawler design and implementation. The task could be constrained by parameters like the maximum number of pages to be fetched (long crawls vs. short crawls) or the available memory. A crawling task can be viewed as a constrained multi-objective search problem.

Genetic algorithms [5] are good at effectively solving large search and optimization problems. They are a group of stochastic search algorithms discovered in 1960s inspired from evolutionary biology. GA searches a space defined by the representation of the application by iterating the three basic step like Fitness Evaluation, Selection and applying genetic operators.

Section 1 gives the introduction to information retrieval and importance of genetic algorithm in information retrieval and crawling. Section 2 gives the related work done by different authors related to term co occurrences. Section 3 includes the proposed architecture of the crawler based document retrieval using genetic algorithm. Section 4 covers the results of the proposed work and Section 5 gives the conclusion.

A. Relevance of Genetic Algorithm in Information Retrieval

There are three main components that have to be taken care while designing GA. 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, population structure and genetic operators.

Genetic algorithm is a powerful search mechanism and it is suitable for the information retrieval for the following reasons [4].

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. GA is more efficient than

using a hill climbing algorithm. The traditional methods of query expansion manipulate each term independent of other. GA contributes to maintain useful information links representing a set of terms indexing the relevant documents.

The traditional methods of relevance feed back are not efficient when no relevant documents are retrieved with the initial query. The probabilistic exploration induced by GA permits the exploration of new areas in the document space.

B. Functions of the Document Crawler

Document Crawler is used especially for searching information such as journal, conference publications from the online databases and in other document collections. User can give reference to the location where all the documents are stored. The Crawler scans each and every documents and it stores the title of the document, document id, the abstract of the document and the keywords given in the document. It extracts all the information content from the document.

II. Related Literatures

In [1] Helen J. Peat and Peter Willett 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 [6] and treating the members of the same class alike for document retrieval. So use of co-occurrence data is having disadvantages.

The indexing of documents and queries is enhanced either by replacing a term by a thesaurus class or by adding a thesaurus class to the index data. The number of documents is much larger than the number of terms in the database. So document classification is much more expensive [7] [3] and has limitations. The term relations are generated on the basis of linguistic knowledge and co occurrence statistics. The similarities between terms are then calculated by using these modifiers from the list based on some syntactic context [8] [2] which has some difficulties. But the queries need not contain similar or co occurring words in it. A query may be formulated with terms that fall within totally dissimilar domains.

For instance, "Metrics for Bioinformatics" Here, metrics and bioinformatics are not co-occurrence terms and hence they cannot be considered to be in the same class. Hence use of co-occurrence data for document retrieval is not considered to be an efficient method. To overcome the limitations, a new efficient method for finding combination of terms is proposed, which helps in a better document retrieval. This method is good for retrieving the documents in which the key words don't have any related meaning.

III. Genetic Algorithm For Document Retrieval With Crawler: Proposed Approach

The proposed architecture is shown in Fig.1

A. Problem Definition

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 is 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.

B. Proposed Approach

The keywords extracted from the document collections are stored in the database. A frequency measure is associated with each keyword. For making search process more efficient, the concept of combining the keywords in the term list is introduced. Combination of the keywords[9] plays an important role in retrieving the relevant information. Here we are using a genetic algorithm approach to obtain the set of the best combination of the keywords. If any two keywords occur within the same frequency, one of the keyword frequencies is boosted by adding 0.1 to it, to make the keyword frequencies unique. These keywords are used to create a best set of the term combinations based on the fitness function. Thus the obtained best combination terms are stored in a combination list. The advantages of the proposed approach save time and retrieves the most relevant document when a query is given.

C. Genetic Algorithm

Generate Initial Population by the frequencies of the keywords

Repeat

Apply the Fitness function, f to each individual.

Apply Random Selection

Apply One Point cross over

Generate new offspring (better combinations keywords)

Until termination condition.

Representation of Chromosomes: The keywords which are repeated more number of times in the document collection are termed as high frequency terms and the keywords which are repeated less number of terms is termed as low frequency words. The mean of the frequency has to be calculated and the value has to be kept as the threshold value and then keywords are grouped accordingly. The keywords in the term list are grouped as high frequency terms and low frequency terms and stored as hfreqterm list and lfreqtermlists respectively.

Documents are scanned by the crawler and the keyword and associated frequency is stored. The sum of all the keyword frequencies stored in the database is found. The mean of frequencies are also found. The frequencies above the mean value are termed as high frequency words and the

frequencies that are less than the mean value are termed as low frequency words. Each term is considered as a gene which shows the frequency of the particular term in the document.

The initial population is represented by randomly picking the term from the high frequency and the low frequency terms. Each keyword has a unique frequency. 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.

Let the keywords and the corresponding frequencies be { grid computing, 4.01; job grouping, 5.02; Active networks, 1.01; Mobile agents 1.30; genetic algorithm,2.04 ;information retrieval,2.06 ; clustering ,1.07; Genetic Programming ,1.29 } and another set of chromosomes {computational grid, 1.09; ecommerce, 1.24; agents, 2.01; grid ,5.01; cosine similarity,1.20 } Fig.2 shows the chromosome representation of each keyword. Each gene in the chromosome shows the frequency associated with each keyword.

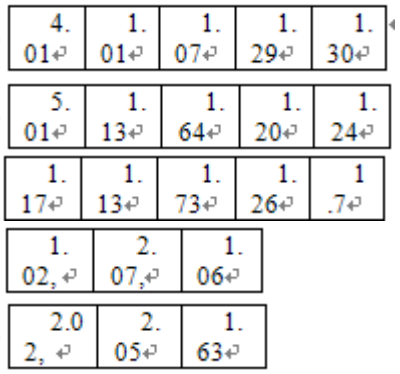

Figure 2. Different Chromosome representations

Fitness Function: The Fitness function used is

Fitness function, $F = 1 - \frac{n}{N}$

where n 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 collection. the fitness values are given below in the table:

For finding fitness of best combination terms, three keywords are considered at a time for finding the fitness functions.

Chromosome	Fitness Function
1.02, 2.07, 1.06	0.89181
1.49, 2.48, 2.09	0.91897
1.45, 3.01, 1.13,	1.0237
2.02, 2.05, 1.63	0.97297
1.39, 1.14, 3.29	1.0853

TABLE 1 Fitness Calculation

The combination term formed by randomly picking three terms is calculated. In the fitness function, we can find the first two combination terms are having more fitness than the third one. So when selection is applied we can ignore the third chromosome. We have to select the best chromosomes from the combination of three terms and two terms based on the fitness function.

Genetic Operators: Tournament Selection is used as the selection operator. The cross over used is single point cross over. If, after generating the new population, the fitness function is no longer improving then terminate the run.

GA Parameters	Value
Population Size	50
No. of Generations	100
Length of the Chromosomes	5
Selection	Tournament Selection
Cross over	Single point cross over

TABLE 2 GA Parameters

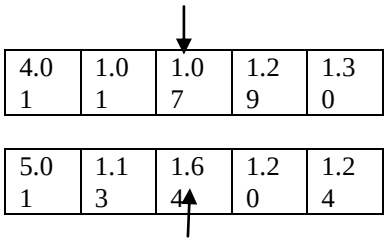

Figure 3. Parents: Before Cross Over

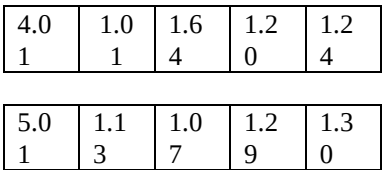

Figure 4. Offspring: After Cross Over

The input to the system is a set of keywords. The output obtained is the set of best combination terms and they represent the possible solutions to the problem. The chromosomes are randomly generated from the high frequency terms and the low frequency terms. Each chromosome is evaluated by a fitness function. This best set of the combination terms is applied in information retrieval system for obtaining the relevant results. Evaluate the information retrieval system with evaluation measures precision and recall. Precision is the fraction of the documents retrieved that are relevant to the user's information need. Recall is the fraction of the document s that is relevant to the query that is successfully retrieved.

Information retrieval system[10] consists of Document Database and Query module as shown in the figure 3.

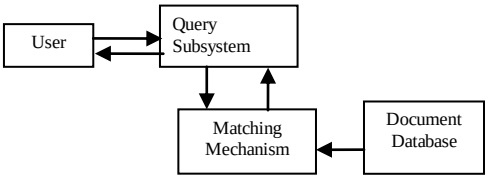

Figure 5. Information Retrieval System framework

Document Database stores the best combination terms which is generated by Genetic Algorithm. The query module processes the user query to find the more relevant documents. Query module consists of two subsystems: Matching Module and Ranking Module. Matching module retrieves the documents which matches the query. It searches in the database in which the combination terms are stored. Ranking module ranks the document according to the relevance of the user query. IR system ranks the documents according to the similarities between document and the query. If a document has got high similarity, then that document is closer to the query .After processing the query effectively, the top most relevant documents are retrieved and it is given to the user.

After deriving the combination terms, the user gives a query to the information retrieval system and it is searched against the document database which has the combination terms. The query is compared against the documents and a similarity measure is calculated to find out whether that particular document is relevant to the query or not. If the document is relevant, it is retrieved. After retrieving the relevant documents from the database, they are sorted and ranked them.

IV. Results and Discussion

The document crawler extracts the information from the journals collected and stores in database. The simulated results are shown in table. This proposal has been tested with 1000 documents and the information retrieval system developed gives average results. The evaluation parameters precision and recall shows satisfactory results.

TABLE 3 Data generated by crawler

id	title	abstract	keywords
1	Agent populated active networks	Abstract Active networking aims at transforming passive data carriers to active, dynamically configurable machines that no only pass data to each other but also perform computation on those data.	Active Networks, Mobile Agents, Security
27	Security-Assured Resource Allocation for Trusted Grid Computing	Abstract: A new trust-based model is developed for optimizing resource allocation in a distributed Grid computing environment. Highly shared resources in a Grid create the insecurity and dependability concerns that hinder distributed applications. process as a multi-objective integer-	Computational grids, distributed supercomputing, resource allocation, ,Inetwork security, trust models, dependable computing
30	Trust Models and NetShield Architecture for Securing Grid Computing	Abstract: Highly shared resources in computing grids make insecurity and privacy abuse major obstacles hindering the grid applications. A scalable grid system demands the allocation and release of resources dynamically in response to the variations in workload, user distribution, security and privacy demands.	Computing grids, dynamic resource allocation, PKI models, Internet security, privacy ,Index Terms: Computing grids, dynamic resource allocation, PKI models, Internet security, privacy ,
4	Data replication and resource allocation on cluster grids	Abstract In data grids, distributed scientific and engineering applications often require access to large amount of data (terabytes or petabytes). Data access time depends on bandwidth, especially in cluster grid. Network bandwidth within a grid cluster is larger than across clusters.	Cluster Grid, Job Scheduling, Data Replication
18	Interval set clustering of web users	Abstract. Data collection and analysis in web mining faces certain unique challenges. Due to a variety of reasonsinherent in web browsing and web logging, the likelihood of bad or incomplete data is higher than conventionalapplications. The analytical techniques in web mining need to accommodate such data. Fuzzy and rough setsprovide the ability to deal with incomplete and approximate information.	clustering, interval sets, Kmeans algorithm, rough sets, unsupervised learning, web mining,Keywords clustering, interval sets, Kmeans algorithm, rough sets, unsupervised learning, web mining,
20	Large-Scale Resource Selection in Grids	Abstract. Grid resource selection requires matching job requirements to avail-able resources. This is a difficult problem when the number of attributes for each resource is large. We present an algorithm that uses the Singular Value Decomposition to encode each resource's properties by a single value. Jobs are matched by using the same encoding to produce a value that can be rapidly compared to those of the resources.	Computational grid, distributed resource allocation, nearest ,Keywords Computational grid, distributed resource allocation, nearest ,neighbor search, multidimensional search, high dimensional data space, Singu-,
6	dynamic trust security with grid integration	Abstract: A new fuzzy-logic trust model is proposed for securing Grid computing across multiple resources sites. We developed a new Grid security scheme, called SARAH. This scheme is supported by encrypted channels among private networks. Instead of relying on using PKI certificates, we suggest to build virtual private networks (VPNs) for Grid trust management through a public network.	Computing Grids, fuzzy logic, virtual private networks, trust management, ,Keywords Computing Grids, fuzzy logic, virtual private networks, trust management, ,resource allocation, linear programming, , distributed Computing ,
19	issues in IR	AbstractThe subject of context has received a great deal of attention in the information retrieval (IR) literatureover the past decade, primarily in studies of information seeking and IR interactions.Recently, attention tocontext in IR has expanded to address new problems in new environments. research.	Context in information retrieval; Interactive information retrieval
34	Using Genetic Algorithm to Improve IR	Abstract—This study investigates the use of genetic algorithms in information retrieval. The method is shown to be applicable to three well-known documents collections, where more relevant documents are presented to users in the genetic modification.	Cosine similarity, Fitness function, Genetic ,Keywords—Cosine similarity, Fitness function, Genetic ,

From Table 3, the required information is extracted in terms of doc id, keywords and frequencies are stored separately for convenience which is shown in Table 4.

TABLE 4 Derived information using crawler

Information		
docId	keywords	frequency
1	Active Networks	1.01
1	Mobile Agents	1.77
1	Security	4.02
10	Grid computing	5.02
10	security	1.65
10	security	4.02
10	trust	4.03
11	Genetic Programming	1.29
12	genetic algorithm	2.04
12	information retrieval	2.06
13	Information retrieval	1.41
13	Ranking function;	1.62
14	genetic algorithms	2.05
14	information retrieval	2.06
14	machine learning	1.71
14	webdocuments	1.13

The obtained results are evaluated using the evaluation measures namely precision and recall, the values of which are shown in Table 5.

TABLE 5 Precision and Recall of proposed system

No of documents	Precision	Recall
50	58.33%	71.4 %
100	60.00%	66.67%
250	60.00%	68.23%
500	58.50%	70.94%

A. Screen Shots

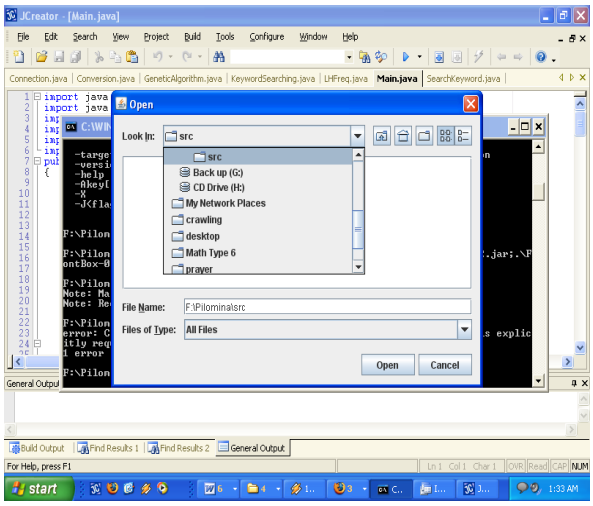

Figure 6. Choosing the document collection from a specified folder

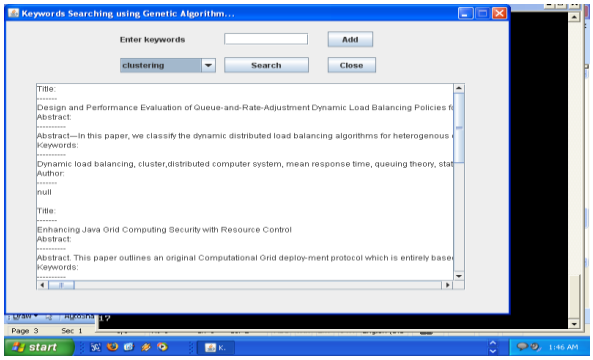

Figure 7. Information Retrieval System

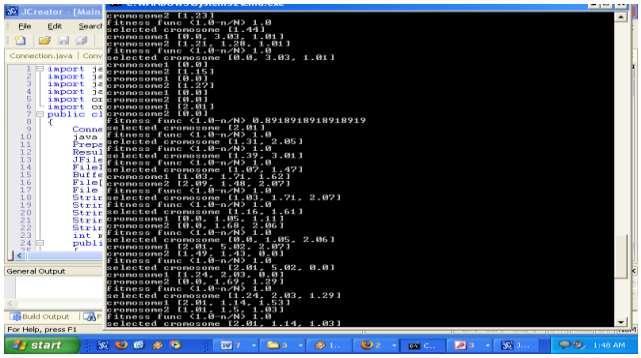

Figure 8. Deriving combination terms using GA

V. CONCLUSION

The proposed information retrieval system is a two stage approach that uses genetic algorithm to obtain the set of best combination of terms in the first stage. Second stage uses the output which is obtained from the first stage to retrieve more relevant results. Thus a novel two stage approach to document retrieval using Genetic Algorithm has been proposed. The proposed information retrieval system is more efficient within a specific domain as it retrieves more relevant results. This has been verified using the evaluation measures, precision and recall.

REFERENCES

[1] Helen J. Peat and Peter Willett*,"The Limitations of Term Cooccurrence Data for Query Expansion in Document Retrieval Systems", Journal of The American Society for Information Science. 42(5),378-383, 1991

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

[3] Abdelmgeid A.Aly, "Enhancing Information Retrieval by using Evolution Strategies" ,Information theories and applications Vol 15, 369-376, 2008

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

[5] David E Goldberg ,Genetic Algorithms in Search, Optimization, Machine Learning , Addison Wesley , 1989

[6] M.E., Lesk, "Word-word association in document retrieval systems",American Documentation, 20(1), 27-38, 1969.

[7] C.J., Crouch, "An approach to the automatic construction of global thesauri", Information Processing & Management, 26(5): 629-40, 1990.

[8] G.Grefenstette, "Use of syntactic context to produce term association lists for retrieval", SIGIR'92, 15th Int.ACM/SIGIR Conf. on R&D in Information Retrieval, Copenhagen , Denmark, 89-97, June 1992 .

[9] Philomina Simon, " A Two Stage approach to Document Retrieval using Genetic Algorithm", International Journal of Recent Trends in

Engineering, Vol 1, No 1, 526-528, 2009
[10] Michael Gordon, Applying probabilistic and genetic algorithms for document retrieval, Computer Practies, 1208 -1218 , 1988

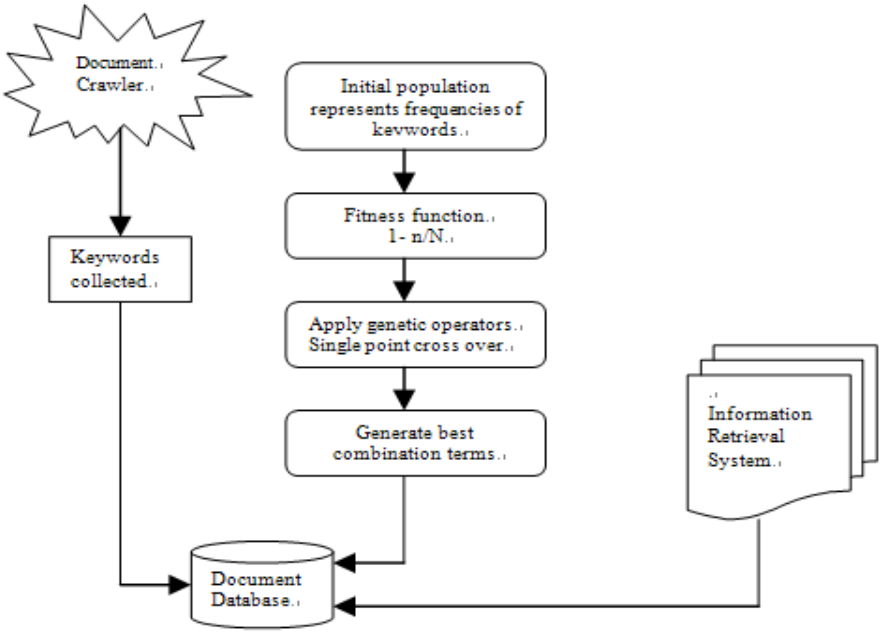

Figure 1. Proposed Architecture: Document retrieval using Genetic Algorithm with Crawler