Information and Knowledge Management
ISSN 2224-5758 (Paper) ISSN 2224-896X (Online)
Vol.3, No.9, 2013

File Search with Query Expansion in a Network System(s)

Noor Ali Ameen Albayaty[*1] Nushwan Yousif Baithoon [1,2]
1. ICCI, Informatics Institute for Postgraduate Studies, Kirkuk, Iraq
2. University of Baghdad ,PO box 47146, Baghdad, Iraq
*E-mail: nurali_albeyety@yahoo.com , nybalnakash@yahoo.com

Abstract

The amount of information in the Internet is growing fast; Searching for information has become an important issue; however the user queries impact the effectiveness of retrieving information that users need. The objective of Query Expansion is to find additional and more relevant results. This article used different similarity measures (Cosine, Jaccard, Dice similarity functions) in VSM on three Genetic Algorithm approaches, each similarity function used as fitness function, one point crossover and new selection method based on rank selection is used. The NSC (National Science Council, Taiwan) document data collection is used in this study. Our results show that QE methods increase the precision rates and the recall rates of information retrieval for dealing with document retrieval . Also we present a network system that consists of many servers to decrease the amount of workload from Main server.

Keywords: information retrieval, vector space model, similarity measures, genetic algorithm, query expansion.

1. INTRODUCTION

The Hypertext-based Webs contain a vast amount of information pertaining to an enormous number of subjects. Therefore, finding useful information pertaining to a particular topic is often a difficult task (Trevor et al., 2001) . In responding to technical information's rapid growth, librarians and information scientists developed the field of information retrieval (Webber) .

Information Retrieval (IR) is the discipline that deals with the retrieval of unstructured data, especially textual documents, in response to a query or topic statement, which may itself be unstructured, a Boolean expression (Greengrass, 2008). The focus of information retrieval is based on 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 (Sathya and Simon, 2009). The goal of IR is to select the informational items (texts, images, videos, etc. which will refer to "documents") that are expected to be relevant for given searcher ("user") from a large collection of such items (Bates, 2012).

Genetic Algorithm GA is a probabilistic algorithm simulating the mechanism of natural selection of living organisms and is often used to solve problems having expensive solutions (Radwan et al., 2006).The especial suitability of GAs to the exploration of very large dimensional vector spaces has led to their being progressively incorporated into AI techniques applied to information science in general, and to information retrieval in particular, since the document spaces deriving from the application of the vector model are real spaces of very large dimensions (López-Pujalte et al., 2003).

2. Classic Models of IR System

2.1 Boolean Model

Boolean model performs a binary indexing in the sense that a term in a document representations either significant (appears at least once in it) or not (it does not appear in it at all). User queries in this model are expressed using a query language that is based on these terms and allows combinations of simple user requirements with the logical operators AND, OR and NOT. The result obtained from the processing of a query is a set of documents that totally match with it, i.e., only two possibilities are considered for each document: to be or not to be relevant for the user's needs, represented by the user query (Hameed et al.).

2.2 Vector Space Model

Represents queries and documents in a high-dimensional space . In this model, documents as well as queries are represented as vectors of weights. Each weight in the vector denotes the importance of the corresponding term respectively in the document or in the query. The vector space is built during the indexing process and contains all the terms that the system encounters (Drias et al., 2009).

2.3 Probabilistic Model

This model tries to use the probability theory to build the search function and its operation mode. The information used to compose the search function is obtained from the distribution of the index terms throughout the collection of documents or a subset of it. This information is used to set the values of some parameters of the search function, which is composed of a set of weights associated to the index terms (Borkar and Patil).

3. EVALUATION OF IR

There are several ways to measure the quality of an IRS, such as the system efficiency and effectiveness, and several subjective aspects related to the user satisfaction. Traditionally, the retrieval effectiveness (usually based on the document relevance with respect to the user's needs) is the most considered. There are different criteria to measure this aspect, with the precision and the recall being the most used (Radwan et al., 2006) .

Precision is a standard IR measure performance .It's defined as a number of relevant documents retrieved divided by the total number of documents retrieved , as shown in eq(1) :

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of retrieved documents in collection}} \quad (1)$$

Recall is a standard IR measure performance .It's defined as a number of relevant documents retrieved divided by the total number of relevant documents in the collection , as shown in eq (2) :

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents in collection}} \quad (2)$$ (King, 2008)

4. QUERY EXPANSION

Query expansion is a process that aims to reformulate a query to improve the results of information retrieval. This is important especially when the original query is short or ambiguous and would therefore give only irrelevant results. By expanding the query with related terms the reformulated query may produce good results (Timonen, 2013).

5. SYSTEM FRAMWORK

5.1 Network System

The proposed Network system consists of two parts: as shown in figure 1:

- Local servers reside at the near end of a user(s).
- Global server resides at the far end of a user(s).

5.1.1 Integrated of Network System:

The information that follows describes the events that will be handled by local and Global server, illustrated in Figure (3.2):-

1. Local server receives a service request from users, it looks for the file using (original query) which will be explanation in section (3.3.1) , If the files does exist, the server forward the file to the requesting user, Else Forward the request to Global server
2. Global server looks for the file using expanded query (Genetic algorithm) which will be explanation in section (3.3.1 and 3.3.2), If the files does exist, the server forward the file to the Local server
3. Local server forwards the file to the requesting user, If the file downloaded by user then store the file in its disk array, Else Return message to the user (File not found)

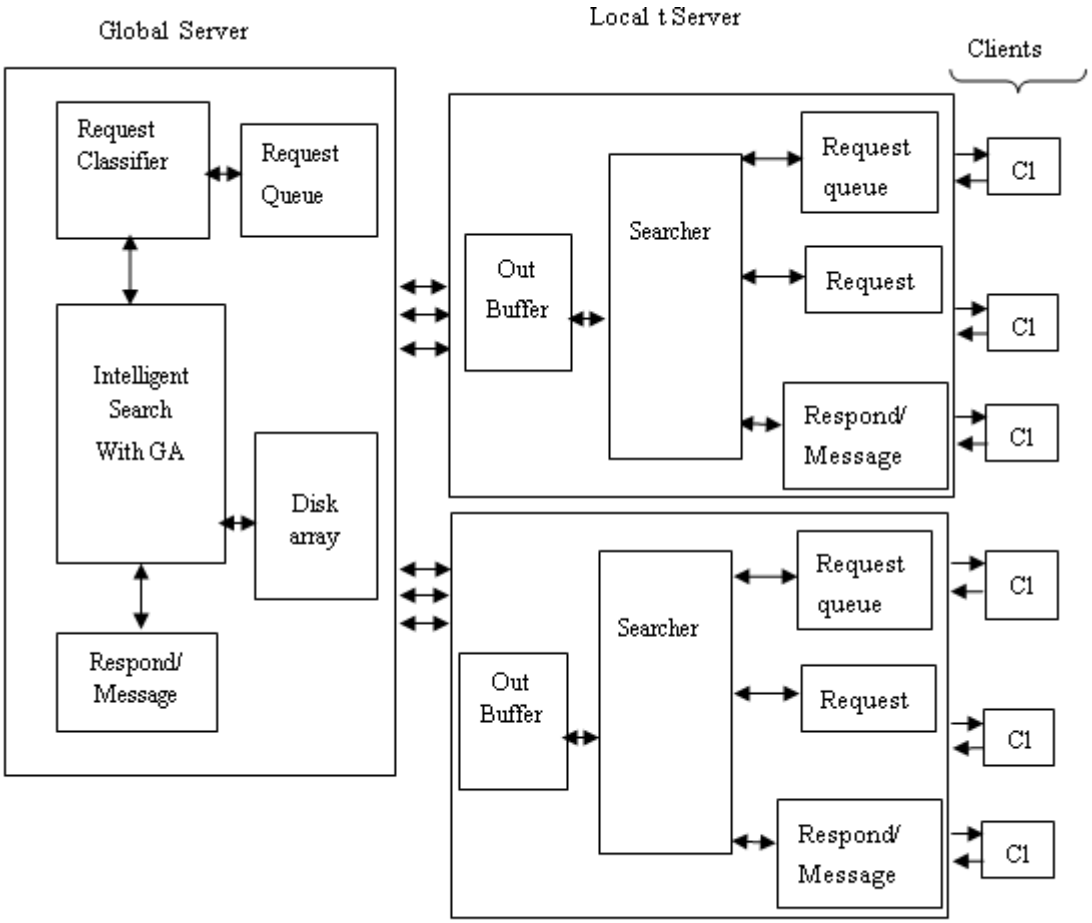

Fig 1: Proposed System Block Diagram

Upon first request in the local region, the local server requests the file from Global Server. Later request of the same file will be in local server and will return directly to users (no request goes to the global server).

5. GA-IR Search System

The Queries are expanded by adding terms that are most relevant to the original query, to be used in the search system, as shown in fig2:

5.1 Building an IR System

The proposed model is based on Vector Space Model (VSM). In VSM, both documents and the user given query are represented as vector of terms.

1. Documents and queries terms are preprocessed, we used the following procedure:

- Create list of keywords from each document titles, keywords "by removing spaces, tabs , new-line characters, and other special characters such as commas ,periods, exclamation points ,and parentheses".

- Elimination stop-words from stop-word list (Kučera and Francis, 1967).

- Stemming the remaining words using the porter stemmer. The most popular stemmer for the English language is the Porter stemmer, as in (Porter, 1980).

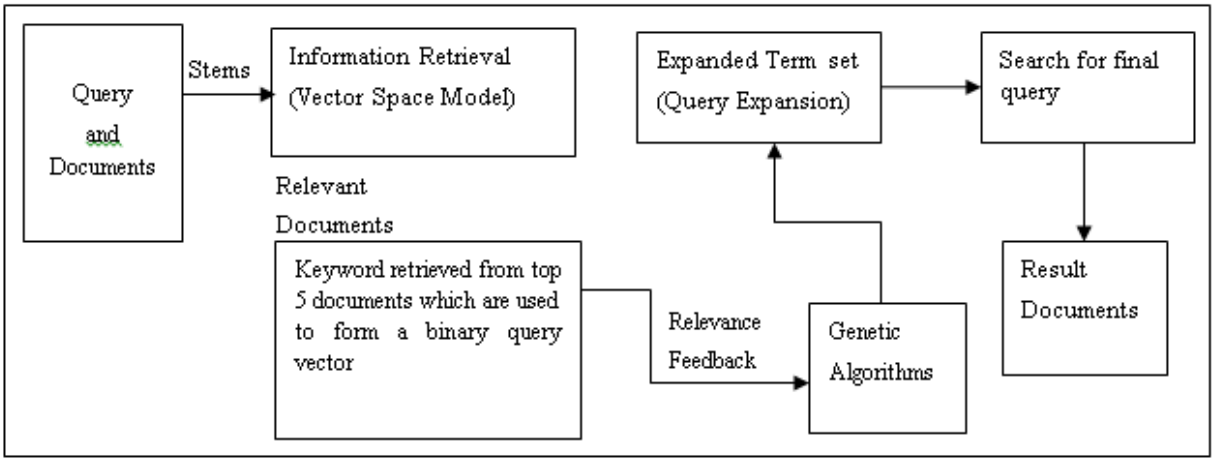

Fig 2: Abstract diagram of proposed intelligent search approach

2. Documents and queries are indexed using Normalized Vector Space Model. The weights assigned to the terms in each document are from the classical tf.idf scheme, The weight of term is measured how often the term j occurs in the document i (the term frequency tfi,j) and in the whole document collection (the document frequency dfj(number of documents containing term j)). The weight of a term j in the document i is:(Jitendra Nath Singh),(Chang and Chen, 2006)

$$w(i,j) = tf(i,j) * idf = tf(i,j) * logD/dfj \qquad (3)$$

Where D is the number of documents in the document collection and Idf stands for inverse document frequency. The normalized frequency of a term j in document i:.

$$f(i,j) = tf(i,j) / max(tf(i,j) * idf) \qquad (4)$$

where f(i,j) =normalized frequency .
tf(i,j) =frequency of term j in documents i.
max(tf(l,j))=maximum frequencies of term j in document i .
The normalized frequency in query Q is given as shown in eq 5:

$$wiq = (0.5 + [0.5 * tf(Q,j) / max(tf(q,j)) * idf] \qquad (5)$$

f(Q,j) =normalized frequency .
tf(Q,j) =frequency of term j in documents i.
max(tf(Q,j))=maximum frequencies of term j in query Q.

3. Documents are matched with queries with Cosine similarity measures
4. Select highest 16 Keywords frequency from the top 5 documents of the list with a corresponding query.
5. These (keywords) are retrieved and then are used to form a binary query vector.
6. The program also generates an output list contains the first results.
7. Adapt the query vector using the genetic approach, to get an optimal or near optimal query vector

5.2 The Genetic algorithm

1. GA approach receives an initial population chromosome corresponding to the keywords retrieved from l IR. The length of chromosome depends on the number of keywords of documents retrieved from a user query

2. The fitness function is a performance measure that is used to evaluate how well each solution is. Given a chromosome, the fitness function must return a numerical value that represents the chromosomes. This score will be used in the selection process. Three different fitness functions are used in the proposed system (Imran and Sharan, 2011)

- The first Genetic algorithm model (GA1) uses a Cosine similarity measure between modified query and chromosomes of the population.

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,q}^2}} \quad (4)$$

- The second Genetic algorithm model (GA2) uses a Dice similarity measure between modified query and chromosomes of the population

$$\text{Dice Similarity} = \frac{2\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,q}^2}} \quad (5)$$

- The third Genetic algorithm model (GA3) uses a Jaccuard similarity measure between modified query and chromosomes of the population

$$\text{Jaccuard Similarity} = \frac{\sum_{j=1}^{n} w_{i,j} w_{i,q}}{\sum_{j=1}^{n} w_{i,j}^2 + \sum_{j=1}^{n} w_{q,j}^2 - \sum_{j=1}^{n} w_{q,j} w_{i,j}} \quad (6)$$

3-Genetic Algorithm module applies the standard GA functions (selection, crossover)

-Selection

The selection process follows the evaluation of the fitness function. The selection is a procedure in which chromosomes are selected for reproduction.

The proposed system first based on Rank Selection which is based on the ranking of the population. The expected selection probability value of each chromosome depends on its rank rather than on its absolute fitness. Because absolute differences in fitness are irrelevant, there is no need to scale fitness (Čepin, 2011). And then new selection method based on rank selection is suggested to reduce the number of terms that will be added to the original query, by making only one crossover among top ranked queries in set of *Query lists* (*QL*), we used the following procedure; select number of chromosomes (n queries) according to the Eq (6): which is depends on a number of generations and the list length.

$$\textit{Query Count} = (\textit{Generation number} - 1) * 2 + (\textit{List Length} - 1) \quad (7)$$

In the beginning there are empty *Best list* and number of *QLs* each one start at increase 2 from precedes it, the last two chromosomes is the result of crossover of the first two in each list, the result of *Best list* is the most efficient queries from comparison among *QLs*, which its words will be used as expanded query .

- Crossover

Crossover is the genetic operators that mix two chromosomes together to form new offspring. The intuition behind crossover is the exploration of new solutions and the exploitation of old solutions . GA's construct a better solution by mixing good characteristics of chromosomes together. We used crossover technique includes one point crossover by choosing a point at random, called crossover point, and exchanging the segments to the right of this point (Klabbankoh and Pinngern, 2000)

Algorithm 1 shows the sequence steps Genetic algorithm based Information Retrieval

Step1: Generate initial population of candidate solution from top ranked chromosome
Step2 : Evaluate initial fitness function of each chromosome
Step3: Set $Initial Generation number ,$Initial list length
Query count = (Generation number-1)*2+(list length-1)
For i=0 to Query count **do**
Generate query list(i)
End for
Step4: Initial Best list =Query list 1
Step5: While Generation number >0 **do**
While Query count >1 **do**
Select first two chromosome from query lis**t**
Perform crossover
Replace it with last two from query list
End while
For i=1 to no of query lists
If query list (i) >Best List
Best list=query list(i)
End for
End while
Step6: Best Query = duplicated words from Best List

6-EXPERMENTAL RESULTS

This paper is dedicated to study the performance of the proposed system was implemented by using Visual Basic 2010, Asp.net.

The subset NSC (National Science Council, Taiwan) document database is used for the experiment. It consists of 520 documents in computer science related fields, split into 28 categories. Each document originally includes title, index number, Chinese abstract and English abstract.

First, the system deletes Chinese abstracts and construct a program to split the document database into files named as its English title and keywords.

To validate the effectiveness of three different methods, and original query, traditional recall and precision are used.

Table 1: User's query terms researches

Queries number	Queries
1	Computer and Music
2	Database Management System
3	Information and Retrieval
4	Intelligent and Agent
5	Mobile and Agent
6	Mobile and Communication
7	Multimedia and Database
8	Network and Secuirty
9	Parallel and Computing
10	Natural Language Processing

Table 2: A comparison of the average precision and the average recall rate for different methods.

	Top 10 Retrieved Documents		Top 20 Retrieved Documents	
	Recall	Precision	Recall	Precision
Original Query	0.364	0.570	0.350	0.345
GA1(Cosine Similarity)	0.469	0.644	0.535	0.520
GA2(Dice Similarity)	0.427	0.632	0.479	0.568
GA3 (Jaccard Similarity)	0.466	0.732	0.483	0.606

From table 2, we can see that the proposed GA methods increase the precision rate and the recall rate of information retrieval system for dealing with document retrieval.

7-CONCLUSION

This paper proposes several contributions , the first contribution is: Query improvement using GA by adding new terms which improve retrieval performance , which uses different types of fitness functions (Cosine ,Jaccard , Dice similarity measures). The second contribution is: building an intelligent search system based on VSM that will use the new queries and compare their results with original query, several experiments have been performed

on NSC document database, that its results showed that these GAs allow to improve the original query. Finally, the third contribution is a network system that consists of many servers one of them is Global Main server (contains all documents) and the others are local servers located at different locations .These servers can communicate with the main server, to decrease the amount of workload from Global server Since later requests of the same file will be in local server and will be closer to users.

References

subset of the collection of the research reports of the National Science Council. Taiwan , R.O.C.http://fuzzylab.et.ntust.edu.tw/NSC_Report_Database/Documents/520documents.html

BATES, M. J. 2012. Understanding Information Retrieval Systems: Management, Types, and Standards, CRC Press.

BORKAR, P. I. & PATIL, A. P. L. H. A MODEL OF HYBRID GENETIC ALGORITHM-PARTICLE SWARM OPTIMIZATION (HGAPSO) BASED QUERY OPTIMIZATION FOR WEB INFORMATION RETRIEVAL.

ČEPIN, M. 2011. Assessment of Power System Reliability: Methods and Applications, Springer.

CHANG, Y.-C. & CHEN, S.-M. 2006. A new query reweighting method for document retrieval based on genetic algorithms. Evolutionary Computation, IEEE Transactions on, 10, 617-622.

DRIAS, H., KHENNAK, I. & BOUKHEDRA, A. A hybrid genetic algorithm for large scale information retrieval. Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on, 2009. IEEE, 842-846.

GREENGRASS, E. 2008. Information retrieval: A survey. 2000. URL citeseer. ist. psu. edu/greengrass00information. html.–fulltext at http://www. csee. umbc. edu/cadip/readings/IR. report, 120600.

HAMEED, S. M., ABDUL-HUSSAIN, M. I. & AHMED, Z. R. RETRIEVING DOCUMENT WITH COMPACT GENETIC ALGORITHM (cGA).

IMRAN, H. & SHARAN, A. 2011. Genetic Algorithm Based Model for Effective Document Retrieval. Intelligent Control and Computer Engineering. Springer.

JITENDRA NATH SINGH, S. K. D. Analysis of Vector Space Model in Information Retrieval". National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC.

KING, J. D. 2008. Search engine content analysis.

KLABBANKOH, B. & PINNGERN, Q. 2000. Applied genetic algorithms in information retrieval. Faculty of Information Technology, King Mongkuts Institute of Techology Ladkrabang.

KUČERA, H. & FRANCIS, W. N. 1967. Computational analysis of present-day American English, Dartmouth Publishing Group.

LóPEZ-PUJALTE, C., GUERRERO-BOTE, V. P. & DE MOYA-ANEGóN, F. 2003. Genetic algorithms in relevance feedback: a second test and new contributions. Information processing & management, 39, 669-687.

PORTER, M. F. 1980. An algorithm for suffix stripping. Program: electronic library and information systems, 14, 130-137.

RADWAN, A. A., LATEF, B. A. A., ALI, A. M. A. & SADEK, O. A. 2006. Using genetic algorithm to improve information retrieval systems. World Academy of Science and Engineering Technology, 17, 6-13.

SATHYA, S. S. & SIMON, P. 2009. Review on Applicability of Genetic Algorithm to Web Search. International Journal of Computer Theory and Engineering, 1, 450-455.

TIMONEN, M. 2013. Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion.

TREVOR, B., WEIPPL, E. & WINIWARTER, W. A Modern Approach to Searching the World Wide Web: Ranking Pages by Inference over Content. INAP, 2001. Citeseer, 316-330.

WEBBER, W. Measurement in information retrieval evaluation. 2010. University of Melbourne, Department of Computer Science and Software Engineering.