International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 4, Number 3 (2014), pp. 239-246

http://www. irphouse.com /ijict.htm

Novel Approach for Query Expansion Using Genetic Algorithm

Pragati Bhatnagar[1] and Narendra Pareek [2]

[1]Department of Computer Science, M.L Sukhadia University, Udaipur, Rajasthan, INDIA.
[2]Department of Computer Science, M.L Sukhadia University Udaipur, Rajasthan, INDIA.

Abstract

This paper is focused towards query expansion, which is an important technique for improving retrieval efficiency of an Information Retrieval System. Specifically the paper proposes a novel evolutionary approach for improving efficiency of Pseudo Relevance Feedback (PRF) Based Query Expansion. In this method the candidate terms for query expansion are selected from an initially retrieved list of documents, ranked on the basis of co-occurrence measure of the terms with the query terms. Top n selected terms create a term pool. From this term pool, Genetic Algorithm is used to select a thematically rich combination of terms, which provide the terms for expanding the query. We call this method as Genetic Algorithm Based Query Expansion (GABQE). The experiments were performed on standard CISI dataset. The results are quite motivating and one can clearly observe the difference in the result when GA is not used and when GA is used. The paper uses GA for improving PRF based query expansion, but at the same time it can also be generalized and tested for other types of query expansions, where terms may be selected in a different way but a good combination of expansion terms can be obtained using GA.

Keywords: Information Retrieval; Query Expansion; Genetic Algorithm.

1. Introduction

Query expansion has been widely investigated as a method for improving the performance of information retrieval system .. Though a lot of work has been done in this area, obtaining a proper expansion of query is still an unsolved problem. A popular type of query expansion is **Pseudo** R**elevance Feedback Based Query Expansion (PRFBQE)**. In PRFBQE, top n documents are retrieved using some efficient similarity measure. Retrieved documents are considered to be relevant. The next step in PRFBQE is selection of expansion terms. Most natural way of selecting the terms is to select the terms which are co-occurring with the query terms. Lesk(1969) expanded the queries by the inclusion of terms that had a similarity with a query term greater than some threshold value of the cosine coefficient. However Rijsbergen(1977) has given the theoretical basis for using co-occurrence statistics to detect the semantic similarity between terms and exploiting it to expand the user's queries. The main problem with the co-occurrence approach was mentioned by Peat(1991), who claim that similar terms identified by co-occurrence tend to occur also very frequently in the collection and therefore, these terms are not good elements to be discriminate between relevant and non-relevant documents. This is true when the co-occurrence analysis is done generally on the whole collection (global) but if we, apply it only on the top ranked documents (local) discrimination does occur to a certain extent. Therefore co-occurrence based techniques have been applied more successfully on PRF based query expansion. However Cao (2008) even question the basic notion of goodness of a term. They argue that a goodness criteria which is based on the frequency of terms in PRF based documents or their distribution in corpus is itself not appropriate. The authors then propose to integrate a term classification process to predict the usefulness of expansion terms The proposed techniques places emphasis on searching for novel material that is related to the search context.

2. Proposed Approach

We have tried to improve the performance of PRF based query expansion by using Genetic algorithm. We call it Genetic Algorithm Based Query Expansion (GAQBE). This helps us to provide thematically rich collection of expansion terms.. GAQBE approach is described below in sec 2.1 and 2.2

2.1 Construction of Term Pool

In order to construct the term pool, we first retrieve top n documents for the query using a matching function which is standard okapi measure here.

All documents and sorted on the basis of standard okapi measure. All the unique terms of top N documents are selected and are ranked on the basis of their co-occurrence with query terms. Top m terms co-occurring with original query terms are selected as candidate terms for expansion. For our experiments, we have used well-known jaccard coefficient as a co-occurrence measure, which is given as :

$$jaccard\_co(t_i, t_j) = \frac{d_{ij}}{d_i + d_j - d_{ij}} \quad (1)$$

Where ti and tj are the terms for which co-occurrence is to be calculated and di and dj are the number of documents in which terms occur, respectively, and dij is the number of documents in which ti and tj co-occur. Inverse document frequency of a term can be used along with above discussed similarity measures to scale down the effect of chance factor. Incorporating inverse document frequency and applying normalization define degree of co-occurrence of a candidate term with a query term as follows:

$$co\_degree(c, t_i) = log_{10}(co(c, t_i) + 1) * \big(idf(c)/ log_{10}(D)\big) \quad (2)$$

$$idf(c) = log_{10}(N/N_c) \quad (3)$$

Where N is number of documents in the corpus, D is number of top ranked documents used ,C iscandidate term listed for query expansion ,tj is jth term of the document, Nc is the number of documents in the corpus that contain c ,co(c,tj) is number of co-occurrences between c and tj in the top ranked documents i.e jaccard_co(c_i, t_j)Above formula can be used for finding similarity of a term c with individual query term. To obtain a value measuring how good c is for whole query Q, we need to combine its degrees of co-occurrence with all individual original query terms t1, t2, t3…. . So we use

$$Suitabilty for Q = f(c, Q) = \prod_{t_i in Q}\big(\delta + co\_degree(c, t_i)\big)^{idf(t_i)} \quad (4)$$

Above equation provides a suitability score for ranking the terms co-occurring with entire Query. The terms of the document are ranked on the basis of similarity value obtained and top m terms form a term pool.

2.2 Genetic Algorithm for Selecting Expansion Terms

Representation of Chromosome-We have used a chromosome representation where each gene represents a specific candidate term. One particular combination of expansion terms represents a chromosome. Considering number of terms as 10, chromosomes are represented in following way

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}

Where each ti represents a term index

Fitness Function-. We have used recall of the retrieved result as a fitness function. Recall is given by :

$$Recall = \frac{|R_a|}{|R|} \quad (6)$$

Ra: Set of relevant documents retrieved and R : Set of all relevant documents

GA operators-Selection ,Crossover and mutation are the GA operators that are applied to the above chromosomes.

Table 1: Algorithm for GABQE.

- Input : Document corpus D
- Query collection Q
- Process:
- Select the query q from Q
- Preprocess the documents in document collection D
- Calculate similarity measure of each document d in D w.r.t query q using equation 1
- Sort documents in D according to their similarity measure with q
- Retrieve top n documents giving document collection R
- Find all unique terms of top n retrieved documents giving term collection T
- Find candidate expansion terms giving term collection C
- calculate co-occurrence between each query term qi and each term ti in T using jaccard similarity
- (equation2)
- Calculate similarity of entire query Q with each term ti in T using equation 3
- Calculate the suitability score of each term ti using equation 5
- Sort the terms in T on the basis of suitability score
- Retrieve top m terms of T giving candidate expansion collection C
- Perform following to select expansion terms by applying GA
- Generate initial population randomly from the term pool.
- Repeat Until the population converges or for maximum number of generation
- Form new population using selection, crossover, mutation operation (in pair of 2)
- Expand the original query by adding terms of the individual population member.
- Retrieve the initial set of documents using tentatively expanded query
- Calculate the fitness of the expanded query using recall based measure (equation 6)
- Return the terms obtained in final generation of GA as final set of expansion terms.
- Output : Set of expansion terms

3. Experiments and Results

We tested the algorithm on CISI dataset. CISI data consist of 1460 abstracts from information retrieval papers and 112 queries. After extensive experiment on corpus the values were set as n = 10 and m = 10.For setting GA parameters chromosome length

was set to 10 as number of expansion terms were 10. Other parameters were fixed after extensive experimentation. The improvement in the result can be observed from recall precision curve shown in Fig. 1 (a)and Fig. 1 (b) .It is observed that GABQE improves the results on average

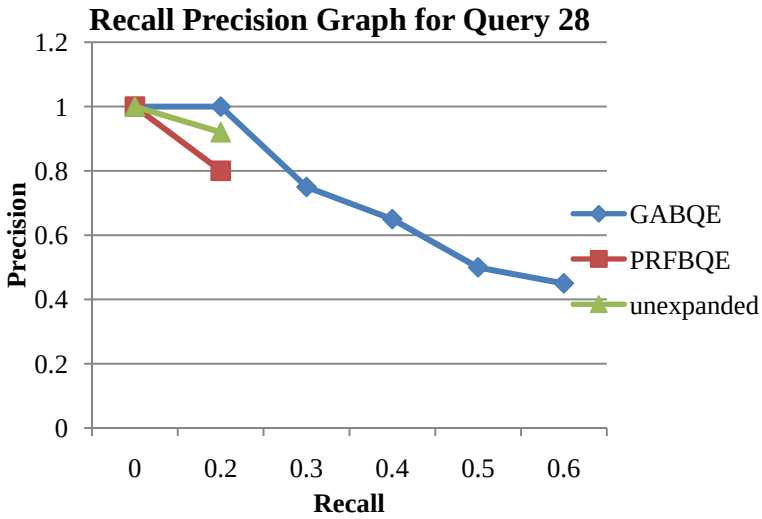

Fig. 1a: Recall Precision Graph for Query 28.

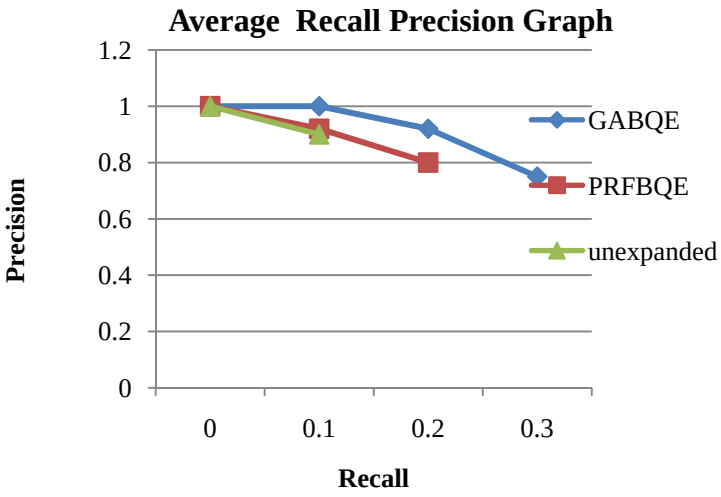

Fig. 1b: Average Recall Precision Graph.

The effect of GABQE can be observed from generation wise average fitness curve. As it can be observed, average recall is increasing and slowly reaches to convergence. This shows that GA is able to improve the fitness (recall), hence efficiency of information retrieval is increased.

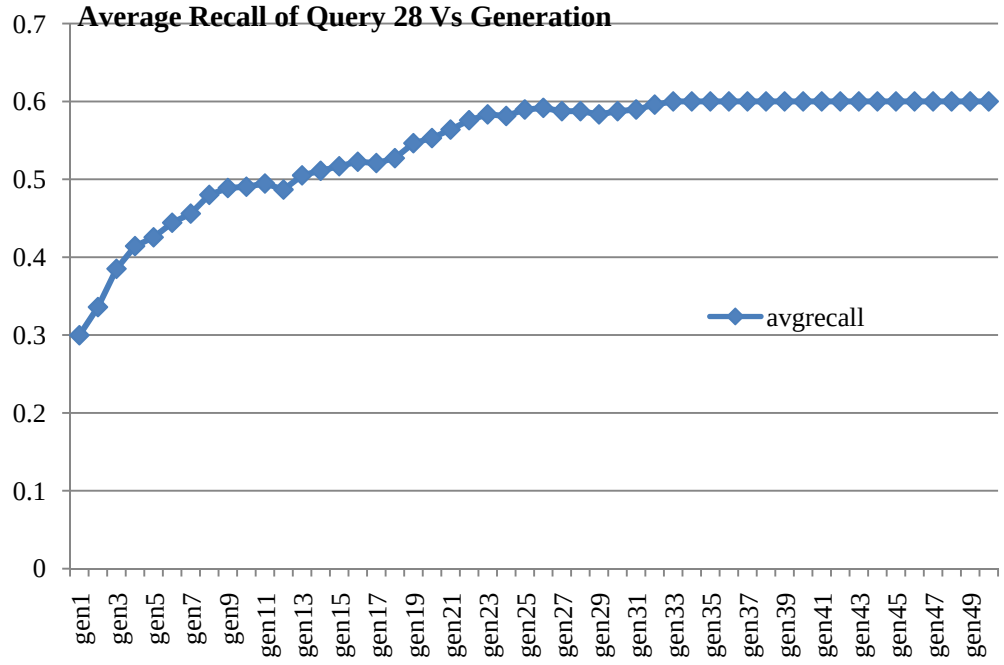

Fig. 2(a). Generation wise Avg. Recall for Query no.28

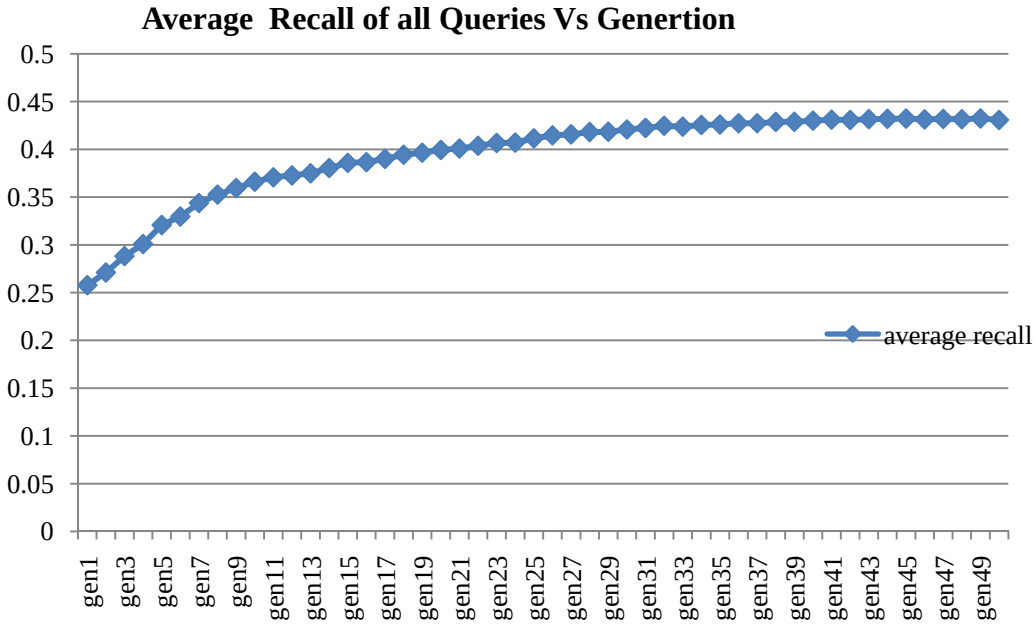

Fig. 2(b): Generation wise Avg. Recall for all Queries

In order to analyze the result we observed the expansion terms obtained without GA and with GA. We observed and analyzed individual expansion terms. In almost all the queries GA based expansion is providing better terms. Table 2 shows recall for original query, for PRFBQE, and for GABQE. In general we can see that GA effects query expansion positively, however the actual effect may vary from query to query.

Table 2: Table showing Recall for PRFBQE and GABQE.

QUERY NO.	RECALL-PRFBQE	RECALL-GABQE
2	(0.377)	**(0.385)**
11	(0.1957)	**(0.2598)**
12	(0.0769)	**(0.6514)**
23	(0.375)	**(0.3958)**
28	(0.2333)	**(0.6)**
(Avg Recall)	0.224	**0.326**

From results it is observed that the terms added before applying GA ie PRFBQE or in the initial run of GA may be relevant individually but in totality, the terms are more cohesive or we can say the expansion terms taken together are more related when GA is applied. So we can say that application of GABQE helps in expanding the query in such a manner that it provides a better selection of thematically rich expansion terms and hence improves retrieval efficiency.

4. Conclusion

This paper suggests use of GA based query expansion (GABQE) in order to improve retrieval efficiency of an Information retrieval system. The experiments have been done on standard CISI collection. The comparison of the result has been done on the basis of recall. The results were compared for unexpanded query, PRF based query expansion and GABQE. It was observed that GA is providing a more cohesive and better selection of expansion terms. The improvement of the result can be observed from the graph. Further we have also analyzed the result by observing the better expansion terms obtained by using our approach.

References

[1] C J Rijsbergen(1977), A theoretical basis for the use of co-occurrence data in information retrieval, J. Documentation, **33**, pp.106–119

[2] G Cao J Y Nie J F Gao S Robertson (2008), Selecting Good Expansion Terms for Pseudo Relevance Feedback,31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 243-250

[3] H J Peat P Willett (1991), The limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems. *JASIS*, **42**, *5*, pp.378–383

[4] M E Lesk(1969), Word-Word Associations in Document Retrieval Systems, American Documentation, **20**, pp.27-38