Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 45 (2015) 52 – 59

Procedia Computer Science

International Conference on Advanced Computing Technologies and Applications

(ICACTA-2015)

An Improved Memetic Algorithm for Web Search

Khushali Deulkar[a]*, Meera Narvekar[b]

[a]*Asst.Prof, Dept. of Computer Engg. D J Sanghvi COE,Vile Parle(w), Mumbai,India*
[b]*Prof, Dept. of Computer Engg. D J Sanghvi COE,Vile Parle(w), Mumbai,India*

Abstract

In order to search a relevant data from World Wide Web, user use to submit query to search engine. Search engine returns combination of relevant and irrelevant results. This paper proposes a novel method based on Memetic Algorithm (MA) for searching the most relevant snippets in case of complex queries. The improved memetic algorithm (IMA) uses a hybrid-selection strategy to enhance the search result. Classical local search operators are combined for improvement in final output. Besides, the same chromosomes are modified to be different so that the population diversity is preserved and the algorithm kept from premature convergence. The performance of IMA is tested by comparing the result of search engine, basic Memetic and Improved Memetic Algorithm. Experimental results show that IMA could obtain superior solutions to the counterparts.

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

*Keywords:*Local Search; Genetic Algorithm; Memetic Algorithm;Snippets.

*Khushali Deulkar . Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address:khushali.deulkar@djsce.ac.in

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).
doi:10.1016/j.procs.2015.03.083

1. Introduction

The internet user uses various search engines to find the information they need. Numerous methods have been developed for searching the most relevant snippets for the fired complex query. Complex query refer to the query with many keyword without forming a proper phrase or sentence. Snippets refer to the data which is retrieved after the query is searched by search engine. The search engine uses various techniques to retrieve the most relevant result. Web spider collects mostly combination of relevant and irrelevant results. Most of the web pages collected by web spider or internet robot are a collection of both relevant and irrelevant results. In proposed system, search engine searches for the snippets for the given user query, these snippets are then processed by IMA to give the most relevant results for the user's query.

There are many methods present to search the data based on clustering. Document clustering[8] algorithms is more efficient in performing the clustering by considering each document as initial centroid and then merges those documents into a cluster by considering the relevancy of contents, until all documents in a cluster have similar feature. Intelligent Cluster Search Engine (ICSE) [9] was proposed based on comparisons of co-occurrences of the term and clustering of documents to give relevant and irrelevant document as an output. Clustering leads to waste of time as filtering of irrelevant information is difficult in this. Different methods for most relevant document search using Genetic algorithm[7] (GA) is proposed over a period of time. Genetic Relation Algorithm[1] is one of the web searching strategies. There is a tremendous scope of improvement in simple Meta heuristic algorithms. Hence the search for new algorithm prevails. Memetic Algorithm[4] (MA) is a combination of evolutionary algorithm (EA) and local search; it applies separate process to refine individuals. It is known that canonical EAs are not well suited to fine tuning search in complex combinatorial spaces but if combined with other approaches can greatly improve the efficiency of search [5]. MA is both more efficient and more effective than traditional EA in many optimization problems [6].

Existing heuristic algorithms for searching in the literature do not pay much attention to the powerful crossover operators, which can also be used to further refine the population. Considering the room for improvement in conventional heuristic algorithm, an improved MA is proposed.

In this paper, the main objective is to find the capability of the MA for searching most relevant snippets. Experimental results show that the proposed algorithm clearly improve the precision of document retrieval when compared with genetic algorithm and memetic algorithm. System architecture is discussed in section 2. The improved memetic algorithm for searching most relevant link is explained in section 3. Computational results and conclusion are discussed in section 4 and in section 5 respectively.

2. System Architecture

In this section the proposed architecture for the system is discussed. The proposed system is based on Memetic Algorithm, which aims at producing most relevant links of web pages (snippets) as the result of web search especially in the case of complex queries.

Here user gives query to the search engine. The result gets stored as a text file into local data base. This raw text file processed by applying cleaning process from which initial population is generated. This data is again reprocessed by algorithm, till the most relevant snippets are obtained.

In our system relationship among various links that are retrieved by conventional search engine will be measured and the most related links will be displayed as output. Fig.1 shows the system architecture for the proposed system.

The main objectives of this system is to get resultant links from the existing search engine, to select better snippets by applying local search algorithm, apply fitness function, mutation and crossover techniques to produce further generations in order to generate most relevant snippets for the complex query.

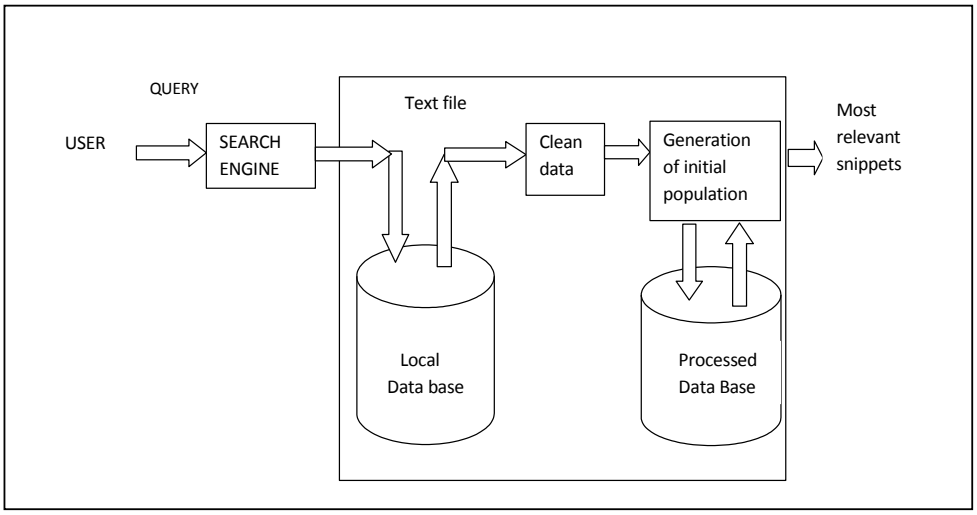

Fig.1. System Architecture

3. Improved Memetic Algorithm

In this section, the Improved Memetic Algorithm is given and how it is different from the other conventional algorithm is discussed in detail. The user first enters the complex query to search engine. The query returns the snippet generated by the search engine. Stemming process is applied to clean both query and snippets. These snippets will be processed to form the initial population for Improved Memetic Algorithm (IMA). The algorithm is given below.

Algorithm for proposed system:
Input: Query string to search
Output: Relevant Pages with links related to query

1. Generate results of conventional search engine(e.g. Google) for the given complex query
2. Copy snippets and store in a text file.
3. Process text file to generate initial population.
4. Perform local search using memetic.
5. The term frequency vector is created using cosine similarity measure.
6. The fitness value of each snippet is calculated.
7. Performs steps of heuristic algorithm(selection, cross over, mutation)
8. Iterate steps 4 to 8.
9. Display final result (the most relevant snippets)

Initial population, chromosomes, preprocessing cosine similarity function, fitness function, type of selection, crossover and mutation which is used for the algorithm is explained in detail below.

3.1. Chromosomes, Initial Population And Pre Processing

Each chromosome is nothing but the snippet which are use to create the initial population. These snippets are then stored in individual text file. These files are then go through cleaning process where stop word, punctuation mark are removed. After pre-processing, the snippets are represented as Term- Frequency vectors, which are then related to each other by means of the cosine similarity measure. A query is defined as set of keywords L.

Row i representing snippet i is a Term Frequency (TF) vector,

$$TFi=\{ TFil1, TFil2, TFil3,..., TFiln\} \quad (1)$$

The quality of snippet i is calculated as,

$$F(i) = \sum_{i \in L} TFil \quad (2)$$

Then Initial population is generated as a weighted matrix of: *No. of snippets x No. of keywords*

Cosine similarity (measure of similarity) between snippet i and snippet j is defined as,

$$S(i,j) = \frac{\sum_{i \in l} TFil * TFjl}{\sum_{i \in l} TFil * \sum_{i \in l} TFjl} \quad (3)$$

Where,

Fil is the frequency if keyword *l* in snippet i.

As an output we get cosine similarity matrix of size n x n, which is used for further evaluation of the snippets.

3.2. Hybrid-selection strategy

The proposed IMA uses the hybrid form of selection. Here the random population is generated with the help of initial population. The chromosomes are same but the order is different in this two. Fig. 2 shows the process of hybrid selection.

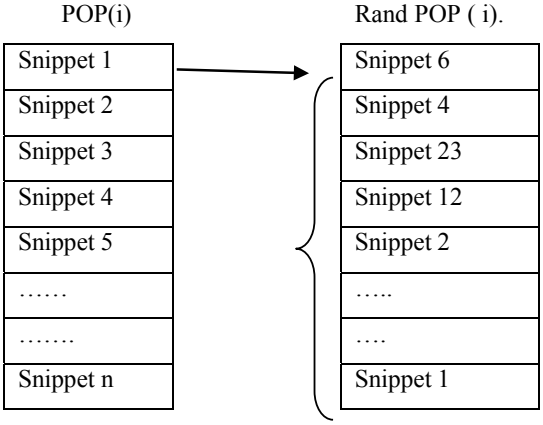

Fig. 2. Hybrid Selection Sort

An index i ranging from 1 to the size of population indicates the hybrid-selection order. The first chromosome from each population is selected. Then, a random number rand between 0 and 1 is generated. A tournament selection is

carried out between POP (i) and Random (i). The best is selected from them. This is hybrid selection.

This kind of selection makes sure that each chromosome is taking part in the competition. In this selection the poor chromosome can also survive if it meets a worse chromosome. The offspring generated by the each cross over are different for the cutting edge chromosome. The fusion of the information of the parents will be better by performing this kind of crossover several times. Generally during crossover process, we only select the newly generated offspring and abandon the parents. This kind of selection may make the offspring worse, but can help to escape from local minima. The selection completely inherits information from the better parent, while the selection process in the crossover operations makes the offspring combine genes from the parents more fully.

3.3. Crossover

Crossover is use to create an offspring for the next generation. It is the process of combining the bits of one chromosome with those of another due to which it inherits traits of both parents. Crossover randomly chooses a point and exchanges the subsequence before and after that point between two chromosomes to create two offspring. Crossover probability means how many couples will be picked for mating.

In the proposed IMA the OX crossover method is used as the crossover operator. The classical crossover operators like Partially Mapped Crossover (PMX), Order Crossover (OX) and Cycle Crossover (CX) can be simply applied to searching[9]. PMX builds an offspring by choosing a subsequence of nodes from one parent and preserving the order and position of as many nodes as possible from the other; OX chooses a subsequence of nodes from one parent and preserving the relative order of nodes from the other; CX preserves the absolute position of the elements in the parent sequence[9]. The relative order of snippets is the key factor of a chromosome. So we choose OX as the crossover operator. The steps of the OX crossover are illustrated in Fig. 3.

First, one cutting point X1 is selected. The first half part of first parent P1 is copied in offspring .The remaining part of offspring is filled by scanning the second half part of the second parent .In this way new child is created by applying the ordered crossover technique. This makes sure the best part of the two parents goes to the next generation.

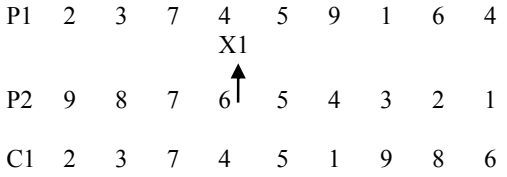

Fig. 3. Ordered Crossover Method

3.4. Mutation

Mutation changes the new offspring by flipping bits from 1 to 0 or from 0 to 1. In order to prevent all solution to fall into local optimum and to preserve diversification in the search, mutation is performed after crossover. Mutation can occur at each bit position in the string with some probability, usually very small (e.g. 0.001). For example, consider the following chromosome with mutation point at position 2:

Chromosome (Before Mutation): 1 0 0 0 1 1 1
Chromosome (After Mutation): 1 1 0 0 1 1 1

A considerable degree of diversity can be obtained by the mutation operator. The mutation operator we applied contains ordered changing mutation where the two points are selected and the bits are swapped within the same chromosome.

3.5. Stopping Criteria

The population can hardly be full of the same chromosome due to diversity preservation method. We can set maximum times of iteration according to the situation of the evolutionary process. Besides, if a current best solution

obtained in the evolutionary process remains unimproved for a specified times of iteration, the MA iteration process can also be stopped to save time.

4. Computational Result

In this section the evaluation parameter, experimental setup and result for the proposed system is discussed.

The experiment was carried out on the platform of personal computer with an Intel®corei3-4010U CPU @1.70GHz. The operating system is Windows 8 and the programs of the mining algorithm are implemented in Java EE (compiled and deployed on Apache Tomcat 6 file server).

The test data set is generated from a JSON file which is prepared using JSON generated script and Google api. When the user submitted a query, search engine returned search results for the user query. Only the top 64 web-snippets were retrieved and displayed to the user. The retrieval of the data is fixed as it depends on execution time of Google api. As most of the users would examine only the top few search results hence this number is enough for the data set. IMA consider the first 64 search result.

The IMA extract the frequently occurring keywords similar to the query entered by user and the term frequency is calculated for every keyword using Equation (1) and fitness value is calculated using Equation (2). After the keyword extraction, relations between these snippets were established using the similarity formula as in Equation (3) without considering the preferences of the user.

4.1 Evaluation parameter

Precision and recall are set based measures to evaluate ranked list. Precision and Recall is mostly used in Information Retrieval domains such as Search Engines. Precision can be plotted against recall after each retrieved document. Precision and recall graph is the most common used method for comparing systems. Typically these graphs slope downwards from left to right enforcing the notion that as more relevant documents are retrieved (recall increases), and the more non-relevant document are retrieved precision decreases.

Precision the ratio of the number of relevant data retrieved to the total number of data retrieved. Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database usually expressed as a percentage[3].

A simple Formula for Precision and Recall:
Precision = Relevant Retrieved /Retrieved
Recall= Relevant Retrieved /Relevant

4.2 Experimental setup

The parameter setups for result analysis is discussed here. The POP _Size depend on the number of snippets extracted. Each crossover generates 64 children and the best one survives. The maximal generation is 64 and the evolution could also be terminated if the global best solution remains unimproved for over 6 iterations.

The test is performed on many complex queries submitted by different categories of user.10 faculties from COMP and IT department of DJSCOE were invited to search complex queries related to their subjects. Then they were asked to mark the generated output as relevant or irrelevant by clicking on the each web- snippets of the results generated by the algorithm and normal search engine(in our case Google).The precision and recall values for each snippets is calculated on basis of how relevant or irrelevant the snippet is. The curve of precision versus recall results by averaging the results obtained. The results are provided in Fig. 4, Fig. 5(a) and Fig.5 (b) respectively.

The algorithms are discriminated by the strategies of selection and local search.

MA : basic selection, with local search procedure.

IMA: hybrid selection, with local search procedure.

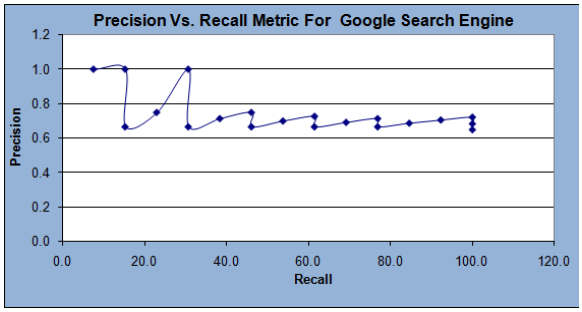

Fig. 4. Result Analysis for Google Search Engine

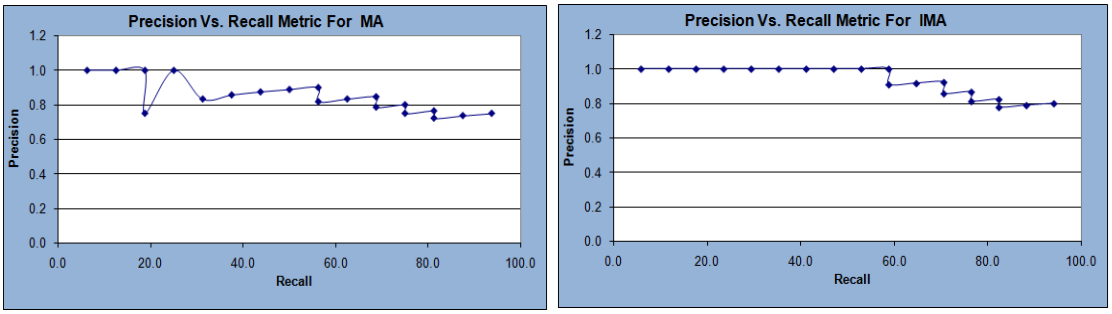

Fig. 5. (a) Result Analysis for MA (b) Result Analysis for IMA

From the above results it is clear that local search with advance improved selection and cross over method may give best result as compared to the result of search engine and heuristic algorithm alone. This approach guarantees to find optimal solutions within reasonable computing time for complex query. The proposed work is a scalable and it can perform well in high performance computing systems.

5. Conclusion

In this paper, an approach for efficient retrieval of relevant snippets has been proposed, in which the similarity between snippets can be compared by considering the co-occurrence terms of a query and snippets. Computational result shows that the proposed algorithm improves the precision of snippet retrieval compared with other conventional heuristic algorithms. The system has a vast scope of future improving as it implemented at experimental level .The improvement may be in terms of reducing search time and number of retrieved links further or by taking the input from many search engines instead of only one. In future, this project can be used as supplement to existing search engines.

References

1. Eloy Gonzales, Shingo Mabu, Karla Taboada,Kotaro Hirasawa “Web Mining using Genetic Relation Algorithm”, SICE Annual Conference 2010 August 18-21, 2010.
2. Poonam Garg “A Comparison between Memetic algorithm and Genetic algorithm for the cryptanalysis of Simplified Data Encryption Standard algorithm” , Institute of Management Technology, India pgarg@imt.edu, International Journal of Network Security & Its Applications (IJNSA), Vol.1, No 1, April 2009.
3. D. D. Lewis, Reuters-21578, distribution 1.0. http://www.Daviddlewis.com/resources/ testcollections/ reuters 21578, 1997.
4. Moscato P. “On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms”. Caltech Concurrent Computation Program, C3P Report. , 826,1989.
5. Krasnogor N, Smith J. A tutorial for competent memetic algorithms: model, taxonomy, and design issues[J]. IEEE Transactions on Evolutionary Computation, 9(5), pp. 474-488,2005.

6. Ong Y S, Lim M H, Zhu N, et al. "Classification Of Adaptive Memetic Algorithms: A Comparative Study"[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 36(1), pp. 141-152, 2006.
7. L. Melita, Gopinath Ganapathy, Sebsibe Hailemariam, "Genetic Algorithms: an Inevitable Solution for Query Optimization in Web Mining – a Review", The 7th International Conference on Computer Science & Education (ICCSE 2012), July 14-17, 2012.
8. M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," Proc. KDD-2000 Workshop on Text Mining, Aug. 2000.
9. M.Sathya, J.Jayanthi, N. Basker, "Link Based K-Means Clustering Algorithm for Information Retrieval", in IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011,978-1-4577-0590-8/11/$26.00 ©2011 IEEE, MIT, Anna University, Chennai. June 3-5, 2011.