Web Information Retrieval Using Genetic Algorithm-Particle Swarm Optimization

Priya I. Borkar and Leena H. Patil

***Abstract*—The rapid growth of web pages available on the Internet recently, searching relevant and up-to-date information has become a crucial issue. Information retrieval is one of the most crucial components in search engines and their optimization would have a great effect on improving the searching efficiency due to dynamic nature of web it becomes harder to find relevant and recent information. That's why more and more people begin to use focused crawler to get information in their special fields today. Conventional search engines use heuristics to determine which web pages are the best match for a given keyword. Earlier results are obtained from a database that is located at their local server to provide fast searching. However, to search for the relevant and related information needed is still difficult and tedious. This paper presents a model of hybrid Genetic Algorithm -Particle Swarm Optimization (HGAPSO) for Web Information Retrieval. Here HGAPSO expands the keywords to produce the new keywords that are related to the user search.**

***Index Terms*—Genetic algorithm, information retrieval system, particle swarm optimization.**

I. Introduction

The most promising information source in the world, the World Wide Web (WWW) is still expanding rapidly. The capacity of storage device is increase and cost is decrease there is tremendous growth in database of all sorts. This explosive growth has led to huge, fragmented and become easy to collect and store information in document collection; it has become increasingly difficult to retrieve relevant information from this large document collection, the search engines play a very important role during this process. Search engines aims to process the enormous information in some collection of document then create an index for quick search. Basically, the index is an inverted file that maps each word in the collection to the set of documents containing that word [1]. Information Retrieval (IR) is a field of study that helps the user to find needed information from a large collection of document. Retrieving information means finding a ranked set of documents that is relevant to the user query [2]. The user with information need issues a query to the retrieval system through the query operational module

Unfortunately, the current commercial information retrieval system that is usually based on the Boolean information retrieval model has provided unsatisfactory results. The GA application is used for information retrieval: What they all have in common is the use of the GA to perform the technique of relevance feedback. In addition, most of the current search engines take up an enormous amount of bandwidth and are time consuming while crawling the web pages [3]. The general objective of information retrieval system is to minimize the overhead can be express at the time a user spend in all of the steps leading to reading an item containing the needed information. The system first extracts keyword from documents and then assigns weights to the keywords, by using the different approaches. Thus, by using the genetic algorithm in this paper presents a model of hybrid GAPSO (HGAPSO) based for effective Web information retrieval. We expand the keywords to produce new keywords that are related to the user search and present more results to users.

II. Genetic Algorithm, Particle Swarm Optimization, Information Retrieval System

A. Genetic Algorithm

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. In GA, the search space is composed of candidate solutions to the problem; each represented by a string is termed as a chromosome. Each chromosome has an objective function value, called fitness. A set of chromosomes together with their associated fitness is called the population. This population, at a given iteration of the genetic algorithm, is called a generation. Genetic algorithms (GAs) are not new to information retrieval [4], [5]. Gordon suggested representing a posting as a chromosome and using genetic algorithms to select well indexes [6]. Yang *et al.* suggested using GAs with user feedback to choose weights for search terms in a query [7]. Morgan and Kilgour suggested an intermediary between the user and IR system employing GAs to choose search terms from a thesaurus and dictionary [8]. Boughanem *et al.* [9], Horng and Yeh [10], and Vrajitoru [11], examine GAs for information retrieval and they suggested new crossover and mutation operators. Information retrieval is one of the most crucial components in search engines and their optimization would have a great effect on improving the searching efficiency due to dynamic nature of web it becomes harder to find relevant and recent information. That's why more and more people begin to use focused crawler to get information in their special fields today. The Fig. 1 Shows the general process of genetic algorithm

Manuscript received November 9, 2012; revised February 23, 2013. A Model of Hybrid Genetic Algorithm-Particle Swarm Optimization (HGAPSO) based Query Optimization for Web Information Retrieval".

The authors are with the Priyadarshini Institute of Engineering and Technology, Nagpur, India (e-mail:priyas1586@yahoo.co.in, harshleena23@rediffmail.com).

DOI: 10.7763/IJFCC.2013.V2.234

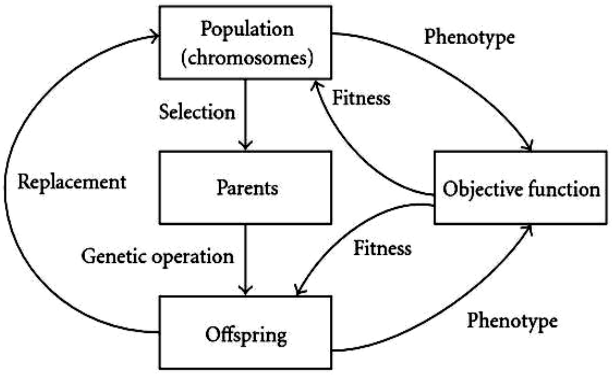

Fig. 1. Genetic algorithm cycle

B. Particle Swarm Optimization

PSO is an evolutionary computation method, which is clearly different from other evolutionary-type methods that does not use the filtering operation (such as crossover and/or mutation) and the members of the whole population are maintained through the search procedure. In order to find an optimal or near-optimal solution to the problem, PSO updates the current generation of particles (each particle is a candidate solution to the problem) using the information about the best solution obtained by each particle and the entire population. Each particle has a set of attributes: current velocity, current position, the best position discovered by the particle so far and, the best position discovered by the particle and its neighbors so far. Each particles start with randomly initialized velocities and positions PSO aims to share information among individuals of a population. In PSO algorithms, search is conducted by using a population of particles, corresponding to individuals as in the case of evolutionary algorithms. Compared to GA, PSO has no operator of natural evolution which is used to generate new solutions for future generation. Instead, PSO is based on the exchange of information between individuals so called particles, of the population, so called swarm. There are two variants of the PSO algorithm were developed, one with a global neighbourhood, and other one with a local neighbourhood. ”In the global neighbourhood, each particle moves towards its best previous position and towards the best particle in the whole swarm, called *gbest* model. On the other hand, according to the local variant, called *lbest* model, each particle moves towards its best previous position and towards the best particle in its restricted neighbourhood” Each particle also adjusts its own position based on its previous experience and towards the best previous position obtained in the swarm. Memorizing its best own position establishes the particles experience implying a local search along with global search emerging from the neighbouring experience or the experience of the whole swarm.

C. Information Retrieval System

Information Retrieval System (IRS), that is, a system used to store items of information that need to be processed, searched and retrieved corresponding to a user’s query. Most IRSs use keywords to retrieve documents. The systems first extract keywords from documents and then assign weights to the keywords by using different approaches [12]. Such a system has two major problems. One is how to extract keywords precisely and the other is how to decide the weight of each keyword.

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 users query. An Information Retrieval System Framework: Three main components of an information retrieval system are shown in Fig. 2. 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 select the relevant document. 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.

A document based IR system typically consists of three main subsystems: document representation, representation of users' requirements (queries), and the algorithms used to match user requirements (queries) with document representations.

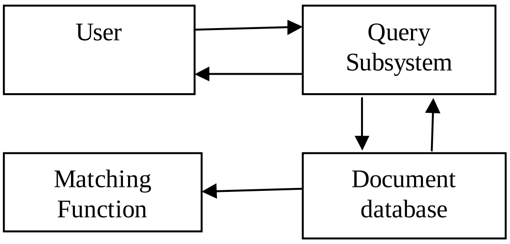

Fig. 2. Information retrieval Framework

1) Components of IRS

An IRS consists of three basic components: Documentary Database, Query Subsystem, and Matching mechanism

- *The documentary database*: This document database stores document along with the representation of their information content. It is associated with the indexer module which automatically generates a representation of each document by extracting the document contents.
- *The Query* Subsystem*:* It allows the user to specify their information needs and presents the relevant documents retrieved by the system to them. The efficiency of an IRS system significantly depends upon query formation.
- *The Matching Mechanism:* It evaluates the degree to which documents are relevant to user query giving a retrieval status value (RSV) for each document. The relevant document is ranked on the basis of this value.

2) Information retrieval models

Boolean model: **-** In the Boolean retrieval model, the indexer module performs a binary indexing in the sense that a term in a document representation is either significant (appears at least once in it) or not. 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.

Vector space model **:-** In this model, a document is viewed as a vector in n-dimensional document space (where n is the number of distinguishing terms used to describe contents of the documents in a collection) and each term represents one dimension in the document space. A query is also treated in the same way and constructed from the terms and weights provided in the user request. Document retrieval is based on the measurement of the similarity between the query and the documents. This means that documents with a higher similarity to the query are judged to be more relevant to it and should be retrieved by the IRS in a higher position in the list of retrieved documents. In This method, the retrieved documents can be orderly presented to the user with respect to their relevance to the query.

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.

III. Hybrid Genetic Algorithm-Particle Swarm Optimization (HGAPSO)

Hybrid Genetic Algorithm-Particle Swarm Optimization (HGAPSO) is proposed by –

Population and Chromosome:- In this paper, the chromosomes from the document are represented directly each document is having weight to represent the weight of the keyword, if weight is zero then the document or keyword is not included in the chromosomes, and the next process of HGAPSO is processed for generating new population.

Fitness function: In HGAPSO, jaccard coefficient is used in the overall process to measure average of the similarity coefficient for each of the training queries against a given document representation. Document representation evolves as described above by genetic operators (e.g. crossover and mutation). Based on weight scheme marks the entire document and select n document at highest marks. Take m words from each document basis of maximum word frequency in a document. Then combine all words of n single document then combine all works of n document and become a single keyword. And convert this word into model 0 and 1 form. Basically, the average similarity coefficient of all queries and all document representations should increase. The jaccard similarity function is finding the fitness value of each document.

Jaccard coefficient:

$$\mathrm{Sin}\ (x, y) = |x \cap y| \div |x \cup y|$$

Genetic operator: - Genetic algorithm operations can be used to generate new and better generations. As shown in Fig. 3 the genetic algorithm operations include:

Reproduction: the selection of the fittest individuals based on the fitness function.

Crossover: is the genetic operator that mixes two chromosomes together to form new offspring. Crossover occurs only with crossover probability Pc. Chromosomes are not subjected to crossover remain unmodified. The intuition behind crossover is exploration of a new solutions and exploitation of old solutions. Gas constructs a better solution by mixture good characteristic of chromosome together. Higher fitness chromosome has an opportunity to be selected more than lower ones, so good solution always alive to the next generation. We use a single point crossover, exchanges the weights of sub-vector between two chromosomes, which are candidate for this process.

Mutation: is the process of randomly altering the genes in a particular chromosome. Mutation involves the modification of the gene values of a solution with some probability. In accordance with changing some bit values of chromosomes give the different breeds. Chromosome may be better or poorer than old chromosome. If they are poorer than old chromosome they are eliminated in selection step. The objective of mutation is restoring lost and exploring variety of data. There are two types of mutation:

Point mutation: in which a single gene is changed.

Chromosomal mutation: where some number of genes is changed completely.

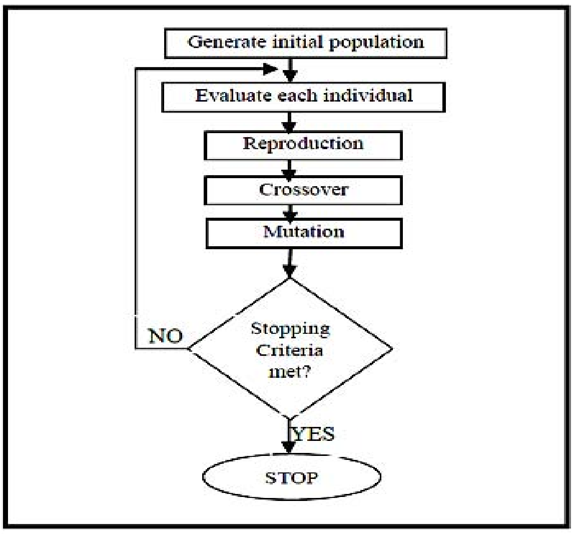

Fig. 3. Flowchart of typical genetic operator

To combine the GA with PSO, the basic elements of PSO algorithm are summarized as follows.

Hybridization is the combination of two or more different things, aimed at achieving a particular objective or goal.

Genetic algorithm and particle swarm optimization are much similar in their inherent parallel characteristics [13]-[17], both algorithms start with a group of randomly generated population; both have a fitness value to evaluate the population.

- P.S.O is one such method where global optimization is done.

- We need an approach which lets us include new particles after initial population selection.
- The approach we have adopted is genetic algorithm in which new population can be included by an operation called mutation if we get trapped in the initial population
- Though we get a better output than other techniques, we need to concretize G.A
- Therefore we adopted a hybrid approach of transitional where one algorithm runs for user defined number of iterations and results obtained passed to the other algorithm alternatively.
- By hybridizing P.S.O and G.A we observe a better output than the individual outputs

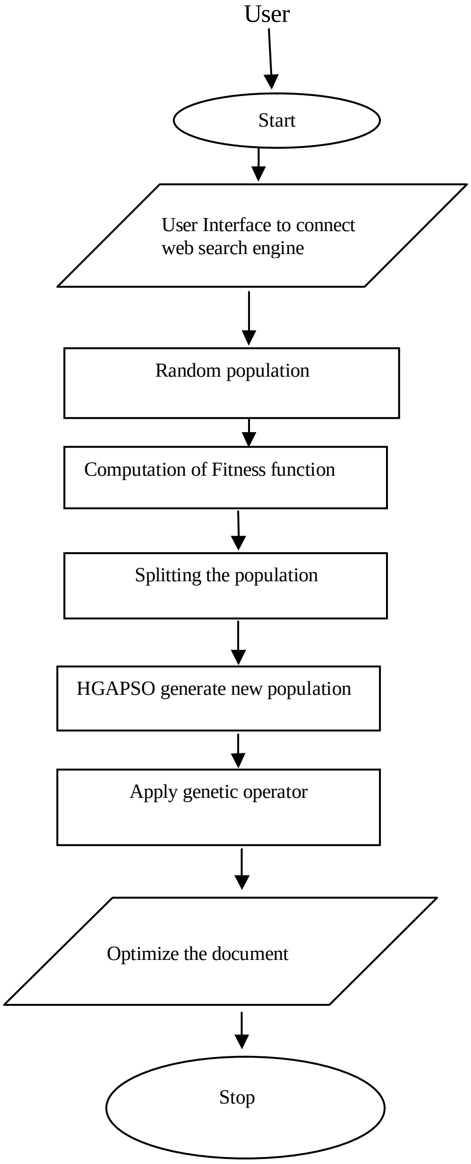
User
Start
User Interface to connect web search engine
Random population
Computation of Fitness function
Splitting the population
HGAPSO generate new population
Apply genetic operator
Optimize the document
Stop

Fig. 4. Data flow diagram of HGAPSO

IV. Data Flow Diagram of Hgapso

Users of the online search engines often find it difficult and tedious to express their need for information in the form of a query. However, if the user can identify examples of the kind of documents that they require then they can employ a technique known as HGAPSO. User searches the query or document then it is generated random population. HGAPSO is applied to find the relevant search pages; it expands the keyword to produce the new keyword. The main advantage of keyword optimization is effective information retrieve using genetic and particle swarm optimization. The primary concern in representation is how to select proper index terms. Query formatting underlying model of retrieval used model. A matching algorithm matches a user's request with the document representation and retrieves document that are most likely to be relevant to the user. The jaccard coefficient fitness function is used to calculate the fitness value. After finding the fitness value the genetic operator is applied to find out the optimize result that represent the relevant document. The data flow diagram of HGAPSO is shown in the Fig. 4. to retrieve the web information in effective way.

TABLE I: Example of User Queries

Queries	Information
Q1	iskandar malaysia development
Q2	iskandar
Q3	iskandar malaysia
Q4	khazanah nasional
Q5	iskandar johor open

V. Experimental Result

In this paper, user will search their interest topic through search system. After user enters the keyword, the system will search the term related to that keyword from the database. Then, the result will be presented to the user. From the interface, a user will select interest topic that is most related to the keyword entered before. After that, the keyword will be arranged in an array to represent the chromosome in binary so that the fitness value for each document can be calculated. Document with high fitness value will be picked in the selection operation. The process flow in our system is listed below:

1) User enters query into the system
2) Match the user query with list of keywords in the database.
3) Encode the documents retrieved by user selected query to chromosomes (initial population).
4) Then, the chromosomes will be processed by the HGAPSO and the new population will be generated.
5) Population feed into genetic operator process such as selection, crossover and mutation.
6) Step 3 is repeated until maximum generation is reached. Then, get an optimize query chromosome for document retrieval.
7) Decode optimize query chromosome to query and retrieve new document from database.

VI. Conclusion

In this paper the evolutionary algorithm help to reformulate a user query to improve the results of the corresponding search. The algorithm uses fitness function which is represented by the equation gives more sophisticated result, a measure of the proximity between the queries terms selected in the considered individual. Then, the top ranked documents are retrieved using these terms

REFERENCES

[1] Z. Zhu, X. Chen, Q. Zhu, and Q. Xie, "A GA-based query optimization method for web information retrieval," *Applied Mathematics and Computation*, vol. 185, no. 2, pp. 919-930 , 2007.

[2] D. Vrajitoru, "Large population or many generations for genetic algorithms? Implications in information retrieval," in F. Crestani and G. Pasi (Eds.), *Soft computing in information retrieval, Techniques and applications*, Physica-Verlag, pp. 199–222, 2000.

[3] S. N. Ibrahim and S. Ali "Query optimization in relevance feedback using hybrid GA-PSO for effective web information retrieval," *IEEE transaction*, 2009.

[4] J. Hyma *et al*., "a new hybridized approach of PSO & GA for document clusters," vol. 2, no. 5, pp. 1221-1226, 2010.

[5] L. Pujalte, C. Bote, and V. P. G. D. M. F. Anegon, "A test of genetic algorithms in relevance feedback. Inf. Process. Manage," vol. 38, no. 6, pp. 793-805, 2002.

[6] P. Bhatnagar and N. K. Pareek, " A Combined Matching Function based Evolutionary Approach for development of Adaptive Information Retrieval System," *IJETAE*, vol. 2, no. 6, June 2012.

[7] A. A. A. Radwan, B. A. A. Latef, "Using Genetic Algorithm to Improve Information Retrieval Systems," *World Academy of Science, Engineering and Technology*, 2006.

[8] C. N. Z. M. Skubacz, "Content Extraction from News Pages Using Particle Swarm Optimization on Linguistic and Structural Features," in *Proc. of IEEE/WIC/ACM International Conference on Web Intelligence*, 2007.

[9] J. S. H. Dom´ınguez and G. T. Pulido, "A Comparison on the Search of Particle Swarm Optimization and Differential Evolution on Multi-Objective Optimization," *IEEE*, 2011.

[10] M. Faheem, E. Sallam, T. Eltobely, and M. Elhamshary, "Rank Aggregation Algorithm using Particle Swarm Optimization for Metasearch Engines," *IEEE*, 2011.

[11] B. Y. Qu, P. N. Suganthan, and S. Das, "A Distance-based Locally Informed Particle Swarm Model for Multi-modal Optimization," *IEEE*, 2011

[12] J. S. H. Dom´ınguez and G. T. Pulido, "A Comparison on the Search of Particle Swarm Optimization and Differential Evolution on Multi-Objective Optimization," *IEEE*, 2011.

[13] R. C. Eberhart, and Y. Shi, "Comparison between Genetic Algorithms and Particle Swarm Optimization," in *Proc. of the 7th international Conference on Evolutionary Programming VII*, vol. 1447. Springer-Verlag, London, pp. 611-616.

[14] X. H. Shi, L. M. Wan, H. P. Lee, X. W. Yang, L. M. Wang, and Y. C. Liang, "An improved genetic algorithm with variable population-size and a PSO-GA based hybrid evolutionary algorithm," in *Proc. of International Conference on Machine Learning and Cybernetics*, vol. 3, pp. 1735-1740 , 2003.

[15] K. Latha and R. Rajaram, "An Efficient LSI based Information Retrieval Framework using Particle swarm optimization and simulated annealing approach," *IEEE Trans. Pattern*, 2008.

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

[17] M. Koorangi and K. Zamanifar, "A distributed agent based web search using a genetic algorithm," *International Journal of Computer Science*, 2011.

Priya I. Borkar is from Priyadarshini Institute of Engineering and Technology College, Nagpur, India. B.E. in Computer Science and Engineering from Priyadarshinhi Institute of Engineering and Technology college, Nagpur, India. Working as a Assistant Professor in Vilasrao Deshmukh college of Engineering, Nagpur, India.