International Journal of Communication Networks and Security

Volume 1 | Issue 4 Article 10

April 2012

SEARCH ENGINES USING EVOLUTIONARY ALGORITHMS

SOWMYA RAVI
S.I.E.S GST, Mumbai, India, somu1510@hotmail.com

NEERAJA G ANESAN
S.I.E.S GST, Mumbai, India, neeraja@gmail.com

VANDITA RAJU
S.I.E.S GST, Mumbai, India, vanditaraju@gmail.com

SEARCH ENGINES USING EVOLUTIONARY ALGORITHMS

SOWMYA RAVI[1], NEERAJA GANESAN[2] & VANDITA RAJU[3]

[1,2&3]S.I.E.S GST, Mumbai, India
E-mail: somu1510@hotmail.com

Abstract – A subset of AI is, **evolutionary algorithm** (EA) which involves evolutionary computation, a generic population-based meta heuristic optimization algorithm. An EA uses some mechanisms inspired by biological evolution: reproduction, mutation, recombination, and selection. A genetic algorithm (GA) is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Working of a **search engine** deals with searching for the indexed pages and referring to the related pages within a very short span of. Search engines commonly work through indexing. The paper deals with how a search engine works and how evolutionary algorithms can be used to develop a search engine that feeds on previous user requests to retrieve "alternative" documents that may not be returned by more conventional search engines.

Keywords *– Artificial Intelligence, Evolutionary Algorithm, Genetic algorithms.*

I. INTRODUCTION

An application of Evolutionary Algorithm is Search Engines. Web mining can make use of evolutionary algorithms for faster and better evaluation of solutions. Artificial Intelligence is used in computational mathematics. The algorithm makes use of previous results. It computes new results based on what it has learned from previous ones (by lateral thinking). Rest of the paper is organized as follows. Section 2 and 3 giving detailed description of background information,section 4 explains about current scnarios followed by its drawbacks, section 7 onwards focuses on need and development of search engines using evolutionary algorithms.

II. BACKGROUND INFORMATION

Evolution: - A gradual process in which something changes into a different and usually more complex or better form. In artificial intelligence, an evolutionary algorithm (EA) is a subset of evolutionary computation, a generic population-based meta-heuristic optimization algorithm.

Meta-heuristic :-In computer science, meta-heuristic designates a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality.

Candidate solutions:- [1] to the optimization problem play the role of individuals in a population, and the fitness function determines the environment within which the solutions "live".

Artificial evolution (AE):-Describes a process involving individual evolutionary algorithms.In computer science, evolutionary computation is a subfield of artificial intelligence (more particularly computational intelligence that involves combinatorial optimization problems.

III. RELATED WORK

Yann Landrinschweitzer et al. introduced the idea lateral thinking in search engines. André L. Vizine et al.put forth the idea of evolving a search engine to retrieve documents from the web.

In this paper we present the analysis of the studies of the aforementioned academicians and talk about the working and advantages of using evolutionary algorithms and genetic algorithms for optimizing the retrieval of web documents.

IV. CURRENT SCENARIO

The advent of e-commerce and corporate intranets has led to the growth of organizational repositories containing large, fragmented, and unstructured document collections. Information retrieval systems, designed for storing, maintaining and searching large-scale sets of unstructured documents, are the subject of intensive investigation. Though it is difficult to retrieve relevant documents from such collections, it is relatively less cumbersome to define categories broadly classifying the information contained in the collection. An information retrieval system, a sophisticated application managing underlying documentary databases, is at the core of every search engine. Fine-tuning the performance of information retrieval systems is essential. One step in optimizing the information retrieval experience is the deployment of Genetic.

Algorithms, a widely used subclass of Evolutionary Algorithms that have proved to be a successful optimization tool in many areas. It improves the retrieval accuracy of search engines that retrieve documents from organizational repositories using a value based approach.

Fitness function:- A particular type of function that prescribes the optimality of a solution (i.e., a

chromosome in a genetic algorithm) so that that particular chromosome may be ranked against all the other chromosomes and optimal ones may be allowed to breed and mix to produce better solutions.

Precision Rate:- In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the search.

Recall Rate:- Recall in Information Retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

V. TECHNOLOGY USED

Genetic algorithm - This is the most popular type of EA. One seeks the solution of a problem in the form of strings of numbers (traditionally binary, although the best representations are usually those that reflect something about the problem being solved), by applying operators such as recombination and mutation (sometimes one, sometimes both). This type of EA is often used in optimization problems.

- Genetic programming - Here the solutions are in the form of computer programs, and their fitness is determined by their ability to solve a computational problem.
- Evolutionary programming - Similar to genetic programming, but the structure of the program is fixed and its numerical parameters are allowed to evolve.
- Evolution strategy - Works with vectors of real numbers as representations of solutions, and typically uses self-adaptive mutation rates.
- Neuro-evolution - Similar to genetic programming but the genomes represent artificial neural networks by describing structure and connection weights. The genome encoding can be direct or indirect.

VI. WORKING OF A SEARCH ENGINE

Early search engines held an index of a few hundred thousand pages and documents, and received maybe one or two thousand inquiries each day.[3] Today, a top search engine will index hundreds of millions of pages, and respond to tens of millions of queries per day.

Web crawling

Before a search engine can tell you where a file or document is, it must be found. To find information on the hundreds omillions of Web pages that exist, a search engine employs special software robots, called **spiders**, to build lists of the words found on Web sites. When a spider is building its lists, the process is called **Web crawling**.[2]

Words occurring in the title, subtitles, **meta tags** and other positions of relative importance were noted for special consideration during a subsequent user search. Other systems, such as AltaVista, go in the other direction, indexing every single word on a page, including "a," "an," "the" and other "insignificant" words.

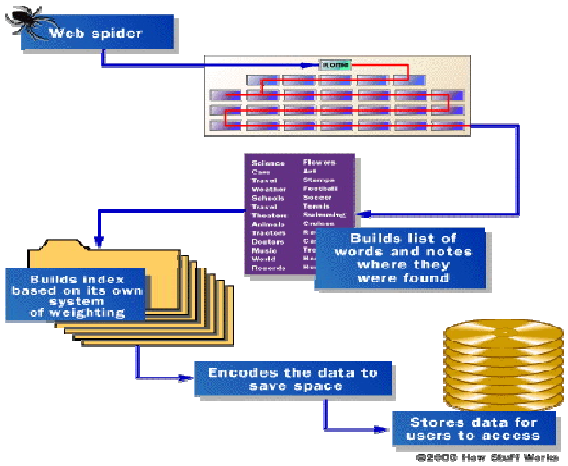

Fig 1: Working of search engine

Meta tags

Meta tags allow the owner of a page to specify key words and concepts under which the page will be indexed. This can be helpful, especially in cases in which the words on the page might have double or triple meanings -- the Meta tags can guide the search engine in choosing which of the several possible meanings for these words is correct. , spiders will correlate meta tags with page content, rejecting the meta tags that don't match the words on the page.

Building the index

Once the spiders have completed the task of finding information on Web pages (and we should note that this is a task that is never actually completed -- the constantly changing nature of the Web means that the spiders are always crawling), the search engine must store the information in a way that makes it useful.

There are two key components involved in making the gathered data accessible to users:

→The **information stored with the data**

→The **method by which the information is indexed**

VII. DRAWACKS OF CURRENT OPTIMAZATION ALGORITHM

Basic search engines that use optimization perform Boolean searches using AND, NOT, OR and NEAR.

The recall rate and precision rate. of search engines that currently use optimization :-

Eg:- A Database where there are 100 documents relating to the general field of "data extraction", a query on "text mining" may retrieve 400 documents. If only 40 of them are about "data extraction" then the tested recall rate will be 40% as the database contains 100 documents on data extraction. As only 40 documents matched the request of the user out a possible 400, the precision rate is 10% only[7].

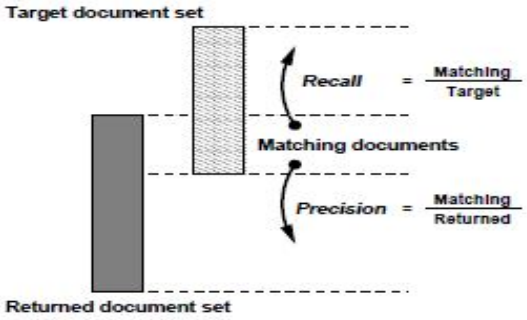

Fig 2: Recall rate and Precision rate

A high precision rate over recall rate is always preferred. That is, it matters more that the result be more relevant than be large in number. It has been proposed that a 'User Profile' can be evolved using genetic programming.

VIII. NEED OF ALGORITHMIC APPROACH

Nowadays, large medical databases consist of a collection of smaller databases, each on different fields and in formats, making it increasingly difficult to retrieve valuable information among the thousands of documents retrieved by a simple query[8]. Databases of large companies grow both linearly and discretely each time another company is absorbed, along with its database. In the end, huge databases of millions of documents and several tera bytes are constituted of several sub databases, each on their own domain, specific structure, language and query systems.

State-of-the-art engines are used to span these databases and they have a high precision and recall rate considering the sheer mass of possibly multilingual data. They feed on previous user requests to retrieve alternative documents that are more relevant, but may not be returned by conventional search engines.

Evolutionary Optimization and application to knowledge extraction:-

It has been observed that a set of 'individual' constitute potential solutions to the problem. At each generation (iteration), new individuals may be created by 'mutation' or 'recombination' of parent individuals. Re-combination explores the genetic pool of the parents while mutation allows new 'genes' to appear. This selection is done to ensure that the best of the individuals are favored/ considered w.r.t to the problem at hand. This process is stochastic as its behavior is non-deterministic, in that a system's subsequent state is determined both by the process's predictable actions and by a random element. The technique, despite being random, is proved to converge theoretically and many successful applications are based on this approach.

Evolutionary algorithmic tools are extremely attractive in case of information retrieval, web mining and document retrieval but are primarily used in off-line implementation, i.e. pre and post processing.

Evaluation step:- A list of documents corresponding to the processed query is presented to the user. The documents actually viewed by the users are considered as interesting and are rewarded accordingly. Modules that rarely contribute to its retrieval of are simply discarded and replaced by newly generated modules. This algorithm allows to evolve a profile that maximizes the "satisfaction" of the user is term of viewed documents.

Information Retrieval:-

Information Filtering :-

A problem that the user has to tackle is "filtering" out relevant information. This is essential due to the sheer mass of data available on the Web. The system has to cater to specific interests of users [9].

Table I – Information filtering × Information retrieval.

Process	Necessary Information	Resources of Information
Information Retrieval	Dynamic	Steady and structured
Information Filtering	Relatively Steady	Dynamic and unstructured

The 3 requisites:-

a) a) Serve specific interests of the user. Irrelevant information must be as little as possible. Number of rejected relevant articles must also be small.
b) The system must adapt to the constant changes of the user.
c) The system should be capable of exploring new domains in order to find some novelties of potential interest to the user.

In systems developed using information filtering based on user profiles, the user specifies which words are of interest and which aren't. If the interest changes, they need to be manually incorporated into the website. This eliminates the need of an explicit feedback by the user on each text.

User Profile:-

Evolutionary algorithms (EA) are used interactively, in order to evolve a "user profile" at each new query. This profile is a set of "modules" that can perform basic re writing tasks on words of the query.

User queries are written with the help of user profiles. DB is searched with the help of re written queries and is presented to the user as a list of documents. User satisfaction is collected as the no. of documents actually read by the user. This is used by the EA as a fitness function, internally. The unread ones are discarded over a period of time. This technique can be used to improve upon Boolean search engines.

Web Mining:-

The internet is the largest library[10]. Its problem being that "Books" are spread all around without indexing. It's the readers' job to find the Web for a site that has his/her desired content. Furthermore the user has to certify the content.

The process of performing information filtering and data mining is termed as web mining[11].

The internet hosts a large number of communities with the most varied interests. Virtual communities allow the aggregation of various societies. An automatic keyword extraction method and Genetic Algorithm (GA) is proposed to search the web. It's designed to be implemented in an academic virtual community. Group profiles will be updated based on the suggestions made by the community.

Keyword Extraction:-

Terminology mining, term extraction, term recognition, glossary extraction or keyword extraction, is a subtask of information extraction. The goal of terminology extraction is to automatically extract relevant terms from a given corpus.

Modeling the growing number of communities and networked enterprises that use the internet, and their information needs is important for several web applications, like topic-driven web crawlers, web services] recommender systems, etc. The development of terminology extraction is essential to the language industry. One of the first steps to model the knowledge domain of a virtual Community is to collect a vocabulary of domain-relevant terms, constituting the linguistic surface manifestation of domain concepts.

IX. GENETIC ALGORITHM

It is a search heuristic that mimics the process of natural evolution and is thus routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as mutation, selection, and crossover.

Mutation: -

The EA periodically makes random alterations in one or more members of the current population, yielding a new candidate solution which may be better than the existing ones.

Crossover: -

The EA attempts to combine elements (decision variable values) of existing solutions in order to create a new solution, with some of the features of each "parent"

Selection: -

The EA performs a selection process in which the 'most-fit' members of the population survive, and the 'least-fit' members are eliminated. The process is the step that guides the EA towards even-better solutions.

X. WORKING

Following describes the working of the proposed algorithm in a virtual community characterized as a scientific paper collection[6]. For users to have access to these sources of information he/she will have to login in 1 or more area of interest. The system autonomously generates group profiles for web documents by selecting a suitable library of keywords, and a search agent that generates and optimizes, via a genetic algorithm (GA), search queries for a search engine. The libraries of the group profiles take into account the relative frequency of a word in a given document and its relative frequency in a set of related and unrelated documents; an approach taken to insert contest information into the system. The search agent uses GA to optimize the search. Keyword extraction and GA were designed to be employed in an academic virtual community in the near future. The GA along with a performance evaluation chart is presented indicating the effectiveness of the technique. Various avenues for investigation are also discussed, along with the future scope.

Keyword Extraction:-

For a word 'w' to represent a group profile; that is, to be selected as a keyword, it has to be a good descriptor of a group and represent a set of documents belonging to the group or folder D and have the following properties:-

1) Be predominant in D when compared with the other words in D.
2) Be predominant in D when compared to its occurrence in all other sets of documents (folders).

The keyword selection method was taken from [12] and works as follows. Let G(w) be the rank of a word 'w'.

$$G(w) = F^{cluster}(w) \text{ X } F^{coll}(w) \quad (1)$$

$F^{cluster}$ - Relates word 'w' with the other words in a given folder

F^{coll} - Relates word 'w' with all other existing folders or groups.

This way $f_j(w)$ corresponds to the number of times word 'w' appears in folder 'j' i.e. the frequency of word in j, then $F_j(w)$ represents the relative frequency of word, defined as :

(2) $$F_j(w) = \frac{f_j(w)}{\sum_v f_j(v)} : \quad v \neq w$$

$0<F_j(w)<1$ and $\sum_v F_j(w) = 1$.

The purpose of this normalization is to consider the importance of the word compared to others in the folder, over the no. of times it appears, as the latter might be deceptive. The relative frequency $F_j(w)$ will play the roleCof $F^{cluster}(w)$. To determine the representation of word 'w' in all folders, $F^{coll}(w)$, the following equation is used:

$$F^{coll}(w) = \frac{F_j(w)}{\sum_i F_i(w)} : i \neq j. \quad (3)$$

The goodness

$$G(w,j) = F_j(w)\frac{F_j(w)}{\sum_i F_i(w)}. \quad (4)$$

Wor[illegible] greater than a pre-specified threshold θ are allowed to enter the library of keywords. This is performed for all words of each document in all folders.

Genetic Algorithm:-

GA introduced by John Holland and extended by David Goldberg are wide applied and highly successful. Steps :

1) Define objective function.

It's a function that's desired to maximize or minimize. The best element id is chosen from the available set of elements.

2) Encode initial population of possible solutions as fixed length binary strings and evaluate chromosomes in initial population using objective function.

3) Create new population (evolutionary search for better solutions):

- Select suitable chromosomes for reproduction (parents).

-Apply crossover operator on parents with respect to crossover probability to produce new chromosomes (known as offspring).

-Apply mutation operator on offspring chromosomes with respect to mutation probability. Add newly constituted chromosomes to new population.

-Until the size of new population is smaller than that of the current go back to step (3).

-Replace current population by new population.

4) Evaluate current population using objective function.

5) Check termination criteria; if not satisfied go back to step 3.

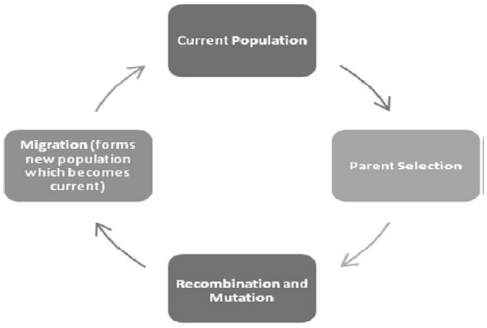

Fig 3:- Iterative process in Genetic Programming

There are 2 populations of chromosomes to be evolved [6]. 1st - each chromosome is composed of a pre-defined number of words randomly defined randomly chosen from the keywords library of each folder. 2nd – contains the same number of genes of the 1st population and same number of individuals of the previous operation. Each gene of the 2nd population may assume 1 of the 3 Boolean random values: AND, OR, NOT.

Ex. illustrating the encoding scheme and query generated by the GA using the chromosomes presented.

Query: results set OR genetic OR selection OR generation -evolution -individual -crossover filetype: PDF

results	set	genetic	selection	generation	evolution	individual	crossover
AND	OR	OR	OR	NOT	NOT	NOT	NOT

AND - default operator. "-" represents the NOT operator. At each generation, the chromosomes of each population are concatenated in order to form a queue encoding query that will be used by the search engine to search for new documents. The document retrieved will be used to determine the fitness of this chromosome for which the cosine measure [13] is used. It determines the similarity between 2 vectors independent of their magnitude. 1 vector represents the library of keywords in the folder, and the other represents the collection of keywords extracted from the document retrieved using the query. The equation returns the angle between these 2 vectors. Its = 1 when the vector points in the same direction and 0 when 90 degrees in angle.

$$sim(D_D, D_Q) = \frac{\sum_{k=1}^{N} W_{Dk} W_{Qk}}{\sqrt{\sum_{k=1}^{N} W_{Dk}^{\ 2} \sum_{k=1}^{N} W_{Qk}^{\ 2}}} \quad (5)$$

W_{DK} - The frequency of the word k in the keywords library of folder D.

W_{QK} - The relative frequency of word k in the document Q.

The 'fitness function' prescribes the optimality of a solution (a chromosome) in a GA so that that a particular chromosome may be ranked against all the other chromosomes. Optimal chromosomes are allowed to breed and mix their datasets producing a new generation that will (hopefully) be even better.

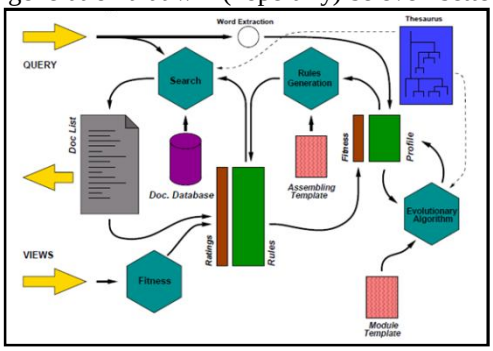

Fig 4 - Working of a Learning Search Engine based on User Profiles [5]

After determining the fitness of all the individuals, a binary tournament is performed to select those individuals that will compose the next generation. The best individual of the population is maintained and is not subjected to crossover (i.e. there is no limit or boundary on it). The crossover operator implemented here was the single-point crossover. The only constraint being that the same word cannot appear twice in the same chromosome in which case, the crossover operator is not applied and the parent chromosomes remain unchanged.

XI. OBSERVATION AND CASE STUDY

This case study presented here has been studied by [6]. When a user (member of the community) finds an article interesting that is still not indexed in the community, this can be suggested for inclusion. With time, these folder structures start to increase in terms of number of documents and in terms of quality of contents, since the papers are selected based on users having common interests and user evaluations. Every time the members of the community suggest a new paper, the group profile will be updated. Group profiles will be composed of a set of keywords from the suggested papers.

Performance Evaluation

To assess the performance, the GA was tested on 3 groups : 1,2,3 using a benchmark database. The keyword extraction process was used to determine the number of keywords from each folder. The chosen threshold was $\theta = 5 \times 10^{-5.}$ The value was chosen empirically and a trade-off was observed between the value of θ and the number and quality of the keywords. High values of θ = few words selected with high goodness value but low vales of θ = high words with low goodness values.

Group	Number of papers	Total of words	Number of words in the group profile
1	347	60,348	36
2	519	75,761	37
3	285	37,876	92

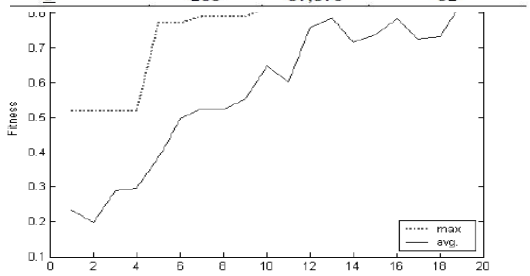

Fig 5: Evolution of the fitness of the best individual (top curve) and fitness of average individual (bottom curve).

GA is capable of improving about 65% quality of the best individual. The diversity of the population at the end of the curve is very less. Mutation may be required.

XII. RELATED WORK

'**GeniMiner**' constitutes of Web Mining with a Genetic-based algorithm.

'**SmartSeek**' is an online querying system that employs a genetic algorithm to adapt to the users' interest. The system accepts user feedback for fitness evaluation.

XIII. DRAWBACKS

The 'Mutator' operation has to be included in the algorithm in order to increase the diversity of the population. EA periodically makes random change in one or more members of the current population, yielding a new candidate solution thereby improving the diversity. In the absence of a 'Mutator' the results might uniform and lack variety. The system is not always capable of retrieving a document for a query. In these cases a new randomly generated query could be introduced such that no individual with a fitness value of 0 is allowed in to the population. The process of randomly generating individuals, improves the diversity.

XIV. CONCLUSION

This paper deals with - Evolutionary algorithms, their types, techniques and relevance to current trends. Operation of most Search Engines, i.e. through optimization. A novel approach towards building Search Engines is using Evolutionary Algorithms. A search engine that works via optimization is relatively easier to implement but there's a possibility of a trade-off between precision value and recall value while processing queries, in quite a few cases. One is always compromised for the sake of the other. Results can be improved up to a certain extent by using "Evolutionary Algorithms" that try to achieve the best of both values by compromising the least

REFERENCES

[1] http://www.geatbx.com/docu/algindex-01.html#P153_5403

[2] http://www.toptut.com/2010/04/11/understanding-web-crawlers-algorithms-for-search-engine-optimization/

[3] http://www.wisegeek.com/how-do-search-engines-ork.html

[4] http://folk.uio.no/nik/2001/08-petrovic.pdf

[5] Yann Landrinschweitzer, Pierre Collet et Al. 'Lateral Thinking in Search Engines using EA'.

[6] Andre Vizin, Leandro Castro et al- Optimizing Web Document Retrieval

[7] D.J. Foskett, "Thesaurus", Readings in Information Retrieval

[8] T. Vachon, N. Grandjean, "Interactive Exploration of Patent Data for Competitive Intelligence".

[9] N.J. Belkin and W. Bruce Croft, "Information Filtering and Information Retrieval: 2 sides of the same Coin?"

[10] J. M. Barrie and D. E. Presti, Colaborative Filtering Science vol. 274, pp.371-372 1996.

[11] R. Cooley, B. Mobasher and J. Srivastava, Web Mining:Information and Pattern Discovery on the WorldWeb Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97),November 1997

[12] K. Lagus and S. Kaski, Keyword selection methodfor characterizing text documents maps. Artificial Neural Networks, (1999), vol 7, pp. 371-376.

[13] G. Salton and M. J. McGill, The SMART and SIREExperimental Retrieval Systems in Readings in InformationRetrieval. Morgan Kaufmann Publishers Inc. (1997),pp. 381-399.