Web Search Query Result Optimization based on Memetic Algorithms: A Comparative Study

Luke Melita[1,2]†, Ganapathy Gopinath [2] and Hailemariam Sebsibe [3]

[1]† School of Computing, Jimma Institute of Technology
Jimma University, Ethiopia

[2] School of Computer Science and Engineering, Bharathidasan University
Tiruchirappalli, India

[3] Department of Computer Science, Addis Ababa University
Addis Ababa, Ethiopia

Abstract

In recent years, the importance and volume of online biological databases, grows tremendously, which makes searching and manipulating data quite challenging.The present study was conducted to minimize the retrieval time of precise data from remote and voluminous databases using metasearch methods. The search query results were optimized by the proposed adaptive cellular Memetic Algorithms (ACMA) implemented over heterogeneous entomological databases. The performance of the algorithm was analyzed by comparing with Genetic Algorithms and with popular search engines like Yahoo, Bing and Ask. The results revealed that ACMA enhanced in efficiency for about thirty-seven times better than the Genetic Algorithms. Despite the growth of data, the performance of the algorithm improves in retrieval time, however making no compromise with its quality.The outcome of the present research could be useful for extending the idea for other similar research areas, and applications like life-science databases with slight modifications.

Keywords: *Adaptive Cellular Memetic Algorithms, Web search query result Optimization, Metasearch engine, Genetic Algorithms*

1. Background

1.1 Databases and Online Databases

Online databases are known for manipulating large sets of data efficiently and retrieving data intelligently from several data sources, and currently they have become an important part of libraries' reference collections too. Online databases also aid in knowledge transmission and promote information exchange between the studies and public audience [1]. Among online databases, biological databases play a pivotal role in assimilating and demonstrating on scientific experiments and in understanding the impact due to biological factors like evolution.

1.2 Significance of data retrieval from Biological and Entomological Databases

Entomological databases and their correlation of information with other biological databases is a part of biocuration that needs greater attention in terms of data management and retrieval, as they determine the health of human and wealth of agriculture. Therefore, it is quite inevitableto bridge the gap between entomological databases andother biological databases. Subsequently, this willenhance the knowledge and indepth understanding in the field of epidemiology, drug development, insecticide, environmental issues, particularly on insectscontrol interventions, population dynamics, bio-diversity, factors associated with density etc. Presently lesser concern has been given for integration of such databases or to have measures for a comparative view or to have a coherent meta-search on relevant databases [2].

At the moment, although only a limited number of entomological databases (EDBs) are available despite their data being enormous than all other living organisms,the available entomological data is sparse and large, with spatial–temporal variation [3]. It is important

to note that several of the existingEDBsare both imprecise and incomplete, however with vast, large and highly variable entomological data [3]. There are volumes of insect information either as independent databases or as museum databases or as single-species collection. Some are structured and online, whereas several others are flat files, documents or excel sheets. Interestingly, numerous publications are available in the databases either for online access or limited with permissions. However efficient retrieval of the heterogeneous information is oftenan obstacle [2]. It becomes still more made complicated, due to sheer volume of data with different data formats and data access methods.

1.3 Why not use a Meta-Search Interface?

Despite the availability of variety of databases, it is quite necessary to identify a gateway to access appropriate information from the right databases, either from a single type of database or in combination from several types of databases. It is mandatory to retrieve more relevant information, in a faster and efficient way. In this context, construction of a meta-search interface/meta search engine to interface with several entomological databases will be an appropriate strategy to retrieve entomological data more effectively within a short-span of time, as it is platform or data model independent. The meta-search engines if built for interfacing with the entomological databases and other biological databases shall have an immense audience and researchers, those surfing frequently for various purposes like disease control, epidemics, drug design, policy decisions and relevant information for further researchin order to enhance the existing knowledge and for the synthesis of new information. Presently either they correlate information manually or use tools to combine a few of the existing databases under a specific entomological society [2].

The meta-search system shall provide facilities to extract information by direct querying for sophisticated users and for building much more precise queries for biological DBs that is user-friendly enough to be used by biologists. The interface shall support queries containing multiple conditions, and shall be able to connect multiple object types, without using the join concept, and shall lessen the burden of the biologists [4]. If the meta-search systems contain intelligent and adaptive interfaces for structured entomological databases, they shall indeed help to obtain fast, effective and optimized results.

1.4 Meta Search Engines – An Overview

Search engines crawl on the web and index the web pages, by which the user search query is responded with results [5]. Metasearch engines are mainly usedto perform a comprehensive search and to extract content from several search enginesefficiently. It is a system that supports unified access to multiple local search engines and databases [6]. They have an innate appeal to improve the retrieval performance ofweb searches.Unlike single source Web search engines, metasearchengines do not crawl on the Internet themselves, instead sendsqueries concurrently to multiple other Web search engines,retrieves the results from each, and then combines theresults from all into a single results listing, avoiding redundancy [7].

The ultimate purpose of a metasearch engine is to diversify the results ofthe queries by utilizing the innate differences of singlesource Web search engines and provide Web searchers withthe highest ranked apt search results from the collection of Websearch engines [7]. However these search engines have been reported to have biases in the arrangement of their results, influenced by various socio-economic and political factors [8-10]. It invariably affects the priority of the web-pages and even more appropriate websites are pushed to the back. Nevertheless the search engines have their specific algorithms to rank the web-pages based on certain parameters like popularity and relevancy.

Several metasearch engines present their results as a single collection from various search engines. However every metasearch engine differs and even contradicts in its functionalities from the other as described here: some combine the results and eliminate duplicates, nevertheless a few display them as separate lists, without eliminating duplicates;

some re-estimates the relevancy of the pages, whereas a few follow the ranking of previous search engines; some store or log the results, where as majority do not store such information; some make some classification on the data, whereas some initiates semantic search; even some are geared towards producing results for a specific topic and in producing visuals on them. Metasearch helps to comprehend information in an easier and faster way than the search engines. Therefore, it calls for constructing a personalized search aggregator that act as a metasearch engineto facilitate users with efficient search results from the web [11].

In general, several sorting, ranking, merging or fusion algorithms are used in the metasearch engines, to improve their efficiency [12-14]. Since the problem of metasearch engine is an optimization problem [15-16], relevant algorithms are observed to be implemented. Among them, bio-inspired algorithms like evolutionary algorithms, particularly genetic algorithms, and particle swarm optimization algorithms, particularly ant colony optimization, etc are the mostly preferred ones [17-22].

1.5 Relevance of the Problem

Due to several huge deductive distributed database systems in the network, the search complexity is constantly increasing and we need better algorithms to speedup query processing. In order to synthesize an appropriateoptimization algorithm, it has been determined that improving memetic algorithms with necessary components will be a solution.

Indeed, in the recent time several researches have been undertaken to resolve the various issuesof meta-search in severalapplication areas. However meta-search on biological data is a complex one, as they require to retrieve data of varyingformats, from a sheer volume of growing data, with different dimensions, and from diverse sources, requiring perfect accuracy of data (approximate data may cause serious errors), however with primitive organization and search capabilities (both in terms of datasource and user).

Though there are a very few query retrieval systems for biological systems like Entrez (under the National Centre for Biotechnology information), Harvester and Ensembl,they mostly deal with the Gene databases. It is surprising to note that the most influential data of 90% of living organisms that is entomology, does not have separate retrieval systems, causing challenges in maintaining and retrieving information on disease pathogens, insect vectors or pests, disease transmission patterns, associated drugs, insect taxonomy, control strategies, statistics and information on their nature and morphology, molecular structure etc. The resultant is retrieval of inappropriate or irrelevant data, mismanagement of data, dispersal of data based on type or region, inaccessibility, and poor-speed and quality of data.

To address these issues, ameta-search engine framed with memetic algorithms is suggested to be implemented based on the entomological databases, to retrieve the related entomological data,andfor other research mappings.

As the performance of the system is critically dependent upon the ability of the query optimization algorithm [23] and the query pre-processing techniques, the chosen computing methodology for query processing, that is memetic computing needs to be improved. So far though genetic algorithms have been used in several cases of search optimization, attention to memetic algorithms for meta-search engine oriented problems has been of lesser concern. Therefore in this work, it has been aimed to improvise the memetic algorithms such that the efficiency and quality of retrieval will be effective.

2. Methodology

2.1 Design and Framework

In order to pursue with this work, analysis on metasearch engines and its operations is initiated. To start with the analysis, different surveys on the available mechanisms were made. While making a detailed review on the relevant literature, meta-search engines are

found to have several common factors and several big areas of research within them. Among them, the query result optimization has been chosen to be the heart of this work. Among several types of ranking measures and their algorithmic techniques, Memetic Algorithms being the most successful among the rest to handle voluminous data with several search parameters, a new framework is proposed. In the constructed framework, the algorithms that need effective procedures or steps are designed, based on the mathematical model.

With result to serious hunt on the big and dispersed data, where meta-search is much needed, biological databases were considered. Later on, pursuing with an intensive study, entomological databases, which is one among the most vital biological databases, however not been concentrated to have effective query retrieval systems, is identified.

Out of the several surveys made, a clear picture on the framework is derived for an Entomological Meta-search interface. The framework is supposed to handle both structured and unstructured data, which may be either from well-organized databases or from unorganized heterogeneous data sources. The query taken as input is parsed and identified whether it needs discrete biological structured data or biological research publications. Then the list of data sources that contain the relevant information alone is passed down with the request. Receiving the retrieved results, the results are optimized based on its content relevance and the ranking provided by the search engines. The feedback on the content relevance and identification of data sources are recorded to make the system adaptive for the next similar or same type of retrieval.

2.2 Mathematical Modeling

It can be seen that web search can be formulated as astandard optimization similar to the problem offunction optimization. Statistical studies have modeled the web as a graph in which the nodes are web pages and theedges are the links that exist between these pages [15, 16].

It is essential to note that the optimization of search, requires to satisfy the both; (1) to optimize the cost of retrieval as well as (2) improving the relevancy of the content. In such a case, the objective function has to incorporate both the functionalities, in terms of a multi-objective function. Therefore the web search being a stochastic search memetic optimization problem has a multi-objective minimization function that depends on both the minimum time taken for retrieval and content relevance of the retrieved result. This is represented in the formula Eq. (1) subject to the constraints that (i) the time taken to retrieve the page together with the associated hyperlinks should not exceed the average time taken and (ii) the distance between the given keyword, and the content available in the page of interest (and the subsequent web-links within the page), should be insignificant. Here the minimum value of the multi-objective function shall give the optimal result. Each of the part of the objective function in formula Eq. (1) can be expanded as specified in formulae Eq. (2) and Eq. (3) that minimizes the time and irrelevancy respectively.

Minimize $Y = f(X) = ((f_1(X), f_2(X))$ **(1)**

Where

$$\textbf{Min } f_1(X) = \sum_{v=0}^{d} R_v(x_j) \ \forall x_j \in X \qquad \textbf{(2)}$$

$$\textbf{Min } f_2(X) = \sum_{p=0}^{ap} T_p(x_j) \ \forall x_j \in X \qquad \textbf{(3)}$$

Satisfying the constraints,

$C_1(X) = \text{lev}(k_i, x_j) \leq 0$ and

$C_2(X) = t_k(x_j) - t_n \leq 0$

Where

$f_1(X)$ → minimizes the time consumed in retrieving the results,

$f_2(X)$ → minimizes the distance between the keyword and the content retrieved,

$C_1(X)$ → checks whether the computed distance is very small and negligible,

lev→Levenshtein's distance function that computes the relevancy of the keyword

$C_2(X)$ → checks whether the time taken to retrieve x_j page is less than the average time taken by another meta-search of the keyword k_i,

X → represents the set of pages and associated hyper-links in the retrieved non-optimized result set and x_j represents the individual

pages
R_v→ represents the content relevance of each type of descriptor in the hyper-text document
T_p→ represents the time required for each process of the algorithm to the retrieve the result set with respect to the keyword k_i
v→ varies from 1 to d, representing the different types of key information within the hypertext document
p→ varies from 1 to ap, representing the time taken to execute each algorithm procedures
i→ varies from 1 to n, representing the keywords and
j→ varies from 1 to m, representing the pages retrieved.

The content relevancy is identified with the help of the Levenshtein's distance formula Eq. (4) [24] that computes the distance between the given keyword, and the page of interest and its associated web-links. The choice of Levenshtein's distance algorithm is due to the retrieval of entomological data where the appropriateness of each keyword search needs to be verbatim. For instance the taxonomy of an insect say species cannot even change by one letter in its spelling.

$$\mathbf{lev_{a,b}(i, j) = \min lev_{a,b}(i-1, j-1) + 1}_{(a_i \neq b_j)} \quad (4)$$

Where a_i→ stands for the given keyword
bj→stands for elements (all headers, title, description, etc) in the web page describing it, which are used for comparison
$1_{(a_i \neq b_j)}$ →is the indicator function that will be equal to 1 if $a_i \neq b_j$ and it will be equal to 0 if $a_i = b_j$.

Similarly the genomic match of an insect, epidemic factors, etc needs exact match of the keyword with the available structured or unstructured entomological data. The Levenshtein's formula for finding the keyword match is given by the formula Eq. (4).

2.3 Memetic Query Result Optimization Algorithm

The retrieved query results are optimized in terms of a memetic algorithm that re-ranks the pages based on their relevance and the priority given by their data sources, so that the resultant order of the pages may not be biased by being away from content relevance and at the same time the priority calculated by the sources based on web crawling is also considered. This will help in eliminating the influential biases.

As discussed earlier in the background section, the efficiency of Genetic Algorithms is shortened by the slow convergence to local optima and rapid diversity loss. However making the memetic algorithms adaptive and cellular, have proved to automatically control local search frequency, where adaptation is carried out basedon the distribution or diversity of a population.In particular, we use fitness values toestimate the level of similarity between individuals and as a basis for applying localsearch on different groups of individuals selectively, thus offering ease of implementation.

The operators involved here are selection, fitness evaluation, reproduction and an extra local search to avoid local hill climbing. This is better described by the following outline in Algorithm 1.

Algorithm 1. Table showing the generic outline of the proposed memetic algorithm

```
Initialize Population P;
Evaluate fitness of P;
While not (stopping condition) do
    P_new:= Select parents;
        Perform Local Search;
    P_r:= Reproduce P_new;   // Crossover or
mutation, to produce an offspring
    Evaluate P_r;
    Output best;
End;
```

The population of retrieved results from all data sources is tested for fitness after which the parents are selected from each of the data sources using the selection operator and is allowed to perform a three-point crossover using the reproduce operator. Mutation occurs when the pages are duplicated. The best among the crossing over candidates is output and the output page is evaluated for fitness once again and collected with the results.

2.4 Implementation Aspects and Method of Execution and Results Collection

Since the system is to measure the efficiency of the query optimization through memetic algorithms over the Entomological databases, a meta-search system that takes from some of the heterogeneous data sources has been constructed. The proposed system shall take entomological data from some of the search engines like Ask, Bing and Yahoo, and from a sample of three structured entomological databases of heterogeneous data modeling types.

The reason why these specific search engines are chosen is because it has been observed that all the major and popular search engines get their data from any one of the following search engines, namely: Google, Ask, Bing and Yahoo. However since Google does not permit external programs to access its information completely, it was necessary to use only the rest. The system is designed in such a way that the time taken by each of the procedures are measured, the retrieval capacities of each of the data sources can be observed, and the relevance of data and their ranks can be listed.

In order to compare with the efficiency of the system, two other systems were developed based on exhaustive search and genetic algorithms respectively. Making several runs on three systems, tabulating and comparing them, shall help to criticize and infer effectively.

The developed system is executed several times with the test data and their results are observed to check with the reliability and stability of the system. Some training data are also fed in and their results are observed.

3. Results and Discussion

3.1 Comparison of Entomological Meta-Search System with other standard algorithms

The memetic algorithms (MAs) designed for optimizing the query results, being implemented as Entomological Meta-search System (EMS) has been run several times (a minimum of ten runs for each case) with different keywords like *Anopheles arabiensis*, butterfly database, entomology database, insect collection etc and also retrieved for different volumes of data.

The time taken to retrieve the results by the EMS is compared with the time taken by the other two systems implemented using Exhaustive Search Algorithms and Genetic Algorithms respectively. The observed time values for running the respective algorithms are tabulated in Table 1.

It can be observed that the mean time values generated are highly comparable with each other, where memetic algorithms (MAs) in EMS prove better than the exhaustive search for about 640 times the running time of MAs and about 37 times better than the Genetic Algorithms. However the standard deviation of the running time is observed to have influenced by the volume of data available in the data sources and the cost required to fetch them. This can be better described by tabulated values of Figure 1, which lists down the average running time of the three listed algorithms (Exhaustive Search, Genetic Algorithms and the Memetic Algorithms) with different quantities of pages retrieved. Figure 1lists down the average running times for 100 and 500 pages retrieval respectively, for the keyword “database”.

It is also necessary to note the difference when the pages are cached and uncached. When the execution is made for the very first time, the

Table 1: Memetic Algorithm Running Time Vs Other Algorithms (in nano seconds)

S.No.	Keyword	Average Time taken by Exhaustive Search Algorithm (ns)	Average Time taken by Genetic Algorithms (ns)	Average Time taken by the Proposed Memetic Algorithms (ns)
1	*Anopheles arabiensis*	11404716	439411	18949
2	butterfly database	26796643	1031491	22193
3	entomology database	12709396	1661364	22601
4	insect collection	6302395	727535	15627
5	Database	17153330	418328	36778
6	Malaria mosquito vectors and disease transmission database	6802305.4	506598.4	17949.4
	Mean time taken (ns)	**14873296**	**855625.8**	**23229.6**
	Standard deviation	7639668.677	482188.9185	7541.692691
	Median	12057056	617066.7	20571

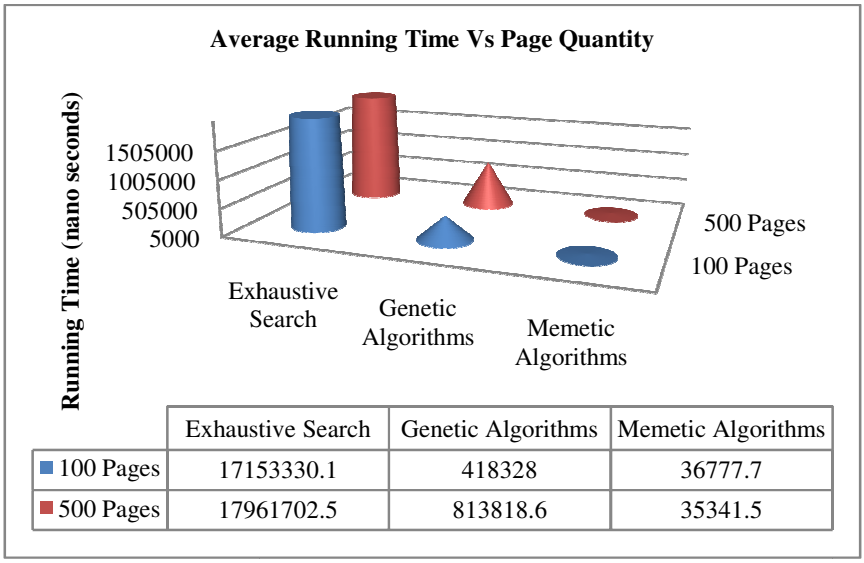

Fig. 1 Average running time w.r.t Page Quantity

Table 2a. Average Time Analysis on Major External activities of the Memetic Algorithm (in ticks per milliseconds)

S.No.	Keyword	Data Fetching (ticks per ms)			Evaluate Page Fitness (ticks per ms)			Logging Candidate Data (ticks per ms)		
		DS1	DS2	DS3	DS1	DS2	DS3	DS1	DS2	DS3
1	Anopheles arabiensis	124487.15	57518.92	80417.43	5866.75	4552.35	4905.81	254.69	142.61	175.19
2	Butterfly database	49423.01	37091.25	66571.68	2643.33	2016.73	4511.44	47.98	98.91	69.73
3	entomology database	75378.90	31675.72	41300.12	3551.87	1892.21	2116.87	97.91	65.57	89.50
4	insect collection	71345.18	81987.09	77419.50	5817.57	4851.69	4418.02	105.39	118.39	237.61
5	Database	17413.47	36081.92	40896.90	745.96	1880.28	2412.66	220.75	261.45	186.30
	Average	67609.54	48870.98	61321.13	3725.10	3038.65	3672.96	145.34	137.39	151.67
	Median	71345.18	37091.25	66571.68	3551.87	2016.73	4418.02	105.39	118.39	175.19
	Standard Deviation	39242.12	21032.49	19166.28	2181.78	1523.06	1302.68	87.99	74.86	70.21

Table 2b. Average Time Analysis on the Major Internal sections of the Memetic Algorithm (in ticks per milliseconds)

S.No.	Keyword	Construct Grid	Selection	Compute Distance	Compute Content Priority	Grid Alignment	Cross Over	Evaluate Generation	Local Search	Collect Best Result	Update Log
1	Anopheles *arabiensis*	0	0.4051	0.0302	24.4125	0.3265	120.214	1.8324	39.9625	0.0184	4.2408
2	Butterfly database	0	0.3886	0.1407	20.1621	0.0001	92.6162	2.0031	30.8719	0.0166	4.3032
3	entomology database	0	0.1851	0.1181	21.758	0.2143	102.234	2.0673	34.0138	0.0249	3.399
4	insect collection	0	0.6154	0.0373	39.6191	0.000031	167.948	1.985	55.9826	0.1871	5.9006
5	Database	0.0079	0.0756	0.201	37.472	0.1929	137.133	2.1081	45.5489	0.1225	3.4666
	Average	0.0016	0.33	0.11	28.68	0.1468	124.03	1.99	41.28	0.07	4.26
	Median	0	0.39	0.12	24.41	0.1929	120.21	2.00	39.96	0.02	4.24
	Std Deviation	0.0035	0.21	0.07	9.16	0.1432	29.90	0.11	9.97	0.08	1.01

time taken to retrieve is comparatively very higher than the rest of the runs made subsequently. This is because of caching of pages by the server, minimizing the running time.

It has been observed that the memetic algorithms do not increase much in running time with increase in page quantity, which makes it suitable for very large databases and huge retrieval from them. Irrespective of the type of execution, that is, whether the data is being cached or not, the average running time of memetic algorithms with different quantities of page retrieval say 100 and 500 pages, are shown respectively in Figure 1.

3.2 Analysis on the efficiency of Memetic Algorithms in terms of running time

In order to analyze the efficiency of the proposed memetic algorithms, the optimal fitness values minimized to prove the objective efficiency is shown in Figure 1, with varying page quantities. It is important to note that though the growth rate of running time increases with page quantity, after a threshold level, the rate of increase in average running time is much lesser or insignificant. This implies that the efficiency of memetic algorithms is not proportional to the increase of voluminous size of data and in fact its efficiency improves with increase in data quantities, lowering the cost of retrieval. The results are quite comparable with the previous study conducted by Garcia *et al* [25], reporting that their Steady-State Memetic Algorithms (SSMA) allows the system to be competitive with other models for Prototype Selection (PS), when the size of the databases increases, tackling the scaling up problem with a good reduction rate and computational time. However SSMA is restricted to make an adhoc local search specifically designed for solving the scaling up problem.

Here in Figure 2, the execution is made for a very lengthy string, say "malaria mosquito vectors and disease transmission database", which requires combination of searches and comparatively higher retrieval cost and query processing, and hence the running time. However, it proves to have a similar growth rate despite its complexity.The success of a system lies with the efficiency of the algorithm in its every part of action. So it is likely to evaluate every major procedure of the algorithm implemented. Table 2a and 2b provide us with the sectional running time of the Entomological Meta-search system in ticks per milli seconds.

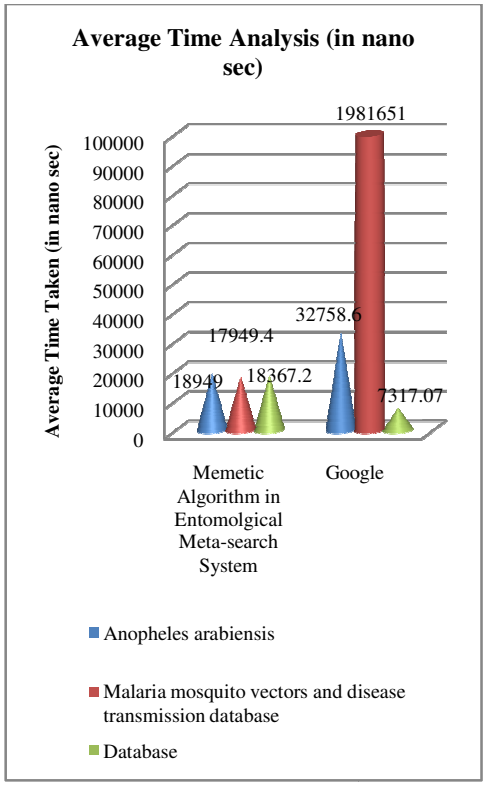

Fig. 2 Comparison of Average time taken – Entomological Meta-search System Vs Google

It is quite obvious from the Table 2a and 2b that the internal activities of the system requires a meager or insignificant amount of time when compared with the external activities like data retrieval from external resources (data fetching), checking URL and logging data for further adaptivity. This strongly affirms that the total retrieval time is greatly influenced by external cost influencers like network speed, traffic associated, request time, response time from the external sources and the database manipulation time within the local server for logging, verifying and updating etc, whereas the algorithm execution time is highly influenced by the quantity and complexity of data.

It can be also observed that the standard deviation of network depended activities are high. This may be due to the fluctuation of network traffic or availability of network resources. However the standard deviation of the internal activities held at the local server is very low as the average does not wander with respect to extremities. Even the median values of internal activities are observed to be closer to their arithmetic mean values maintaining their central tendencies to be normal.

The proposed algorithm proves its efficiency in terms of its adaptivity with the help of logging, cellular arrangement and the local search performed and this is well demonstrated with the help of their average values in the Table 2a and2b. However their mean time takes about 480 ticks per milliseconds (summed up from initial log imaging, update log, grid manipulation, local search and conventional GA processes), which is 430 times smaller than the average total time required for overall algorithm execution. This means that the special nature of Adaptive Cellular Memetic Algorithms (that has been enhanced from popular Genetic Algorithms (GAs)) consumes an insignificant amount of time to improve its efficiency in several folds than the customary GAs. The pattern of the results are similar to the earlier study results reported by Sanusi *et al* [26] that Memetic Algorithm converges faster than Genetic Algorithm even as it also produces more optimal results than Genetic Algorithm produces by a factor of 4.9% when the results obtained from Roulette Wheel selection were compared for both algorithms. The results are also comparable with the findings of Garg [27] indicating that memetic algorithm is an extremely powerful technique for the cryptanalysis of Simplified Data Encryption Standard Algorithm, and that for a very large amount of cipher text the memetic algorithm can be seen to outperform genetic algorithm in terms of efficiency and also in computation time.

3.3 Average Time Comparison with popular Meta-Search Engines

The efficiency of the Entomological Metasearch System implemented in terms of memetic algorithm is compared with some of the popular meta-search engines. Basically several of the successful search engines are based on anyone of the following search engines in retrieving their results directly from them: Google, Yahoo, Bing and Ask, and also they are the top rankers according to Alexa traffic ranking. For these reasons the efficiency of our system is compared with their respective performances.

But as the search engine "Google" forbids testing its efficiency in terms of a program (i.e. controlling through a program is not possible), a manual retrieval is made and the time taken for respectively for specific number of records is observed and hence calculated in nano seconds for retrieving 100 pages at a time and the results are shown in Figure 2.

It can be noted in Figure 2 that the Entomological Meta-Search System shows an average running time for the specified keywords say, *Anopheles arabiensis*, Malaria mosquito vectors and disease transmission database, and database. The keyword "*Anopheles arabiensis*" is very specific to the area of Entomology where the genus *Anopheles* has nearly 430 identified species. This means that this search may involve several combinations of species with the genus Anopheles, and there may be several types of data among the search results (say text files or hypertext documents or structured data etc). All these may increase the complexity of the search. For this keyword, Google takes about 32758.6 nano seconds (in average), which is very high when compared with the time taken for the keyword "database" (7317.07 nano seconds). This is because of the usage frequency of the keyword being indexed with the degree of match or due to the simplicity of the search keyword [28].

However it should be noted that for a very lengthy and complex keyword say "Malaria mosquito vectors and disease transmission database", Google takes enormous time in average for a comparatively very smaller quantity of data (say 1,090,000 page links) (infact the time taken by Google in nano seconds has been approximated to 100 page

links). This clearly shows that the efficiency of a meta-search engine is highly dependent not only on the usage frequency, but also on the complexity and length of the keywords used. This is evidently supported by the earlier study conducted by Wu and Li [29] demonstrating that most of the search services do very well with short queries including only a few key words, however, when added with more words into the queries, the performances go down. Besides, that the order of the words in the query affects the performance of the query result significantly when they differ from the sequence of words in the retrieved documents.

Other than Google, the rest of the search engines say Ask, Bing and Yahoo do not provide the time taken to retrieve the results

Table 3. Retrieval Time Comparison of Search Engines with Entomological Meta-search System (EMS)

S.No.	Keyword	Ask Average Retrieval Time	Bing Average Retrieval Time	Yahoo Average Retrieval Time	EMS Average Retrieval Time
1	insect collection	42332	62247	57388	38309
2	entomology database	71758	32838	37436	40998
3	Entomology	34350	30702	33476	32475
4	Anopheles *arabiensis*	34223	22031	33558	31024
5	butterfly database	63356	42536	59114	56273
	Mean	49203.8	38070.8	44194.4	39815.8
	Median	42332	32838	37436	38309
	Standard deviation	17329.6	15357.3	12945.6	10070.9

and an external program is constructed to measure their retrieval time. Accordingly Table 3 shows the average time taken to retrieve the results from the respective datasources and also from the Entomological Meta-search system (EMS). It is evidential from Table 3 that EMS proves better than the Ask and Yahoo search engines, however equivalent to the efficiency of Bing search engine in terms of running time (although the content relevance of Bing is observed to be lesser than EMS in Figure 3). It can also be observed that the retrieval time depends upon the type and complexity of the keyword and also upon the frequency of usage.

3.4 Content relevance of Entomological Meta-search System

The effectiveness of Entomological Meta-search System is determined by two factors namely: (1) speedy retrieval and (2) effective content. The latter one is proved to be effective from the content of resultant pages retrieved, that has been tested by the program for content relevance and additionally also by manual checking.

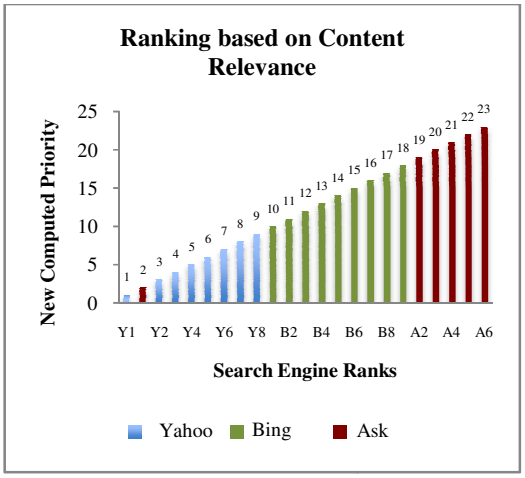

Fig. 3 Content Relevance of EMS vs Popular Search Engines

It is interesting necessary to note that the results ranked are based upon new content relevancy calculation with the help of Levenshtein's distance formula [24]. The new computed distance gives a new rank among the candidates of the same datasource (within yahoo or ask or bing). Then the ranks are used to compare with the rank of all the datasources.

It is performed in the following manner. The new computed rank plus the datasource rank are considered. The highest ranking pages from each data source is taken in order so that the candidates of higher (and equal) ranks are allowed to cross-over. The best among them is output and in order to fill with another parent, a candidate is chosen from the datasource from where the previous best candidate is taken from. This kind of calculation, enables both local search within the datasources and a global search in comparison with the other datasources. Therefore the results in Figure 3 shows both the ranks provided by the data sources and the new implemented system EMS.

Interestingly, EMS not only improves with the efficiency but also in terms of effectiveness by means of checking the content relevance of the retrieved result pages. This has overcome the idea and the results obtained by Oladele [30] demonstrating that the relative performance of GA and MA when investigated for multi-objective optimization of network design, MA outperformed GA in effectiveness while GA outperformed MA in efficiency and that the difference between the effectiveness of MA and that of GA increases as the network's size increases.

Figure 3 compares the ranks produced by the data sources and the new priority computed by EMS. That is, since the non-structured data is retrieved from external sources like Yahoo, Bing and Ask, it is necessary to compare the rankings provided by those search engines with respect to the EMS's ranking. It has been observed that the content relevancy of EMS is better than the other search engines. Among the search engines, Yahoo produces comparatively better results in terms of content relevance than Bing and Ask.

Figure 3 clearly shows how much the results are deviated from the rank of the sources with respect to content relevance. It may be possibly explained that this is due to the popularity of the web pages than their content relevancy or due to the bias of the search engine due to the influence or deposits paid by the web sites, or due to the discrepancy in its estimation of content relevance. Actually a sample execution on the keyword "insect collection" is shown in Table 3.

Figure 3 also exhibits a sample page response from the respective data sources, when expected to retrieve a maximum of ten results from each data source. However only 77% of pages (23/30) have been observed to have responded. It can be observed that the pages' response is affected by the inappropriate citation or addressing, restricted accessibility, inappropriate data, URL transformations/ mistakes or less response from its respective server.

This has been manually tested by the author for each sample (atleast for first 50 top ranking pages output by each of the search engines), by manually visiting or observing the pages and their content. This is also comparable with the studies of [8] and [9] that indicate that search engines have been reported to have biases in the arrangement of their results, influenced by politics, economy and other social issues.

4. Conclusion

This paper after reviewing the need of metasearch on entomological data has made its focus on improving the genetic algorithms for metasearching by introducing local search, adaptivity, restricted mode of search among the neighbourhood and the cellular grid of storage. Moreover the success rate (very slow convergence) of these algorithms is highly influenced by the inclusion of a few factors like query logs and grid manipulation. The results clearly affirm that the system implemented with specified algorithm proves even better than the efficiency of popular search engines, when the search is intuitive and specific with entomology and related biological information. The metasearch system implemented here is not only suitable for entomological datasources but also applicable to any other domain or fields like news, games etc, those having similar approaches. The present research could be easily extended to other several life-science related databases too, with slight inclusions or modifications of guidance parameter tables.

Acknowledgements

Thanking Almighty first, the authors would also like to acknowledge Dr. K. Karunamoorthi for his immense help in editing the manuscript. Our last but not the least heartfelt thanks go to our colleagues from the School of Computing, Jimma Institute of Technology, Jimma University, Jimma, Ethiopia for their kind support and cooperation.

References

[1] S.S. Ningthoujam,A.D. Talukdar, K.S. Potsangbam, M.D. Choudhury (2012) Challenges in developing medicinal plant databases for sharing ethnopharmacological knowledge. J Ethnopharmacol141(1):9-32. doi:10.1016/j.jep.2012.02.042.

[2] L. Melita, G. Gopinath, K. Karunamoorthi, H.M. Sebsibe (2014) Entomological Databases: Challenges and Opportunities in Data Management and Retrieval. International journal of Entomological Research. Vol 2. Issue No. 3 (2014).

[3] S.F. Rumisha, T. Smith, S. Abdulla, H. Masanja, P. Vounatsou (2014). Modelling heterogeneity in malaria transmission using large sparse spatio-temporal entomological data. Glob Health Action.

[4] M. Latendresse, P.D Karp (2010) An advanced web query interface for biological databases. Database 2010. baq006doi: 10.1093/database/baq006.

[5] Jawadekar, S. Waman (2011) Chapter 8. Knowledge Management: Tools and Technology. *Knowledge Management: Text & Cases*, New Delhi: Tata McGraw-Hill Education Private Ltd. p. 278, ISBN 978-0-07-07-0086-4. Available at: http://books.google.com.et/books?id=XmGx4J9daUMC&pg=PA278&dq=%22search+engine+operates%22&hl=en&sa=X&ei=a-

[6] Z. Wu, W. Meng , C. Yu , Z. Li (2001). Towards a highly-scalable and effective metasearch engine. Proceedings of the 10th international conference on World Wide Web, p.386-395, May 01-05, 2001, Hong Kong [doi>10.1145/371920.372093].

[7] B.J. Jansen, A. Spink, S. Koshman (2007) Web Searcher Interaction with the Dogpile.com Metasearch Engine. J Am SocInformaSciTechnol 58(5):744-755. DOI: 10.1002/Asi.20555.

[8] E. Egev (2010) Google and the Digital Divide: The Biases of Online Knowledge, Oxford: Chandos Publishing.

[9] L. Vaughan, T. Mike (2004) Search engine coverage bias: evidence and possible causes. Inform Process Manage 40(4):693-707. doi:10.1016/S0306-4573(03)00063-3

[10] A.D. Corni`ere and T. Greg (2014). Integration and search engine bias. RAND Journal of Economics, Vol. 45, No. 3, Fall 2014, pp. 576–597

[11] L. Melita, G. Gopinath, K. Karunamoorthi, H.M. Sebsibe (2013) Metasearch Engines for Entomological Databases: Existing Challenges and Future Perspectives. In Proceedings of the Fourth Annual Research Conference on Meeting National Development Challenges through Science, Technology and Innovations, Feb 7-8, 2013, Jimma University, Jimma, Ethiopia. pp. 61-66.

[12] C. Liu, Z. Zhang, X. Xie, T. Liang (2008) Evaluation of Meta-Search Engine Merge Algorithms. In Proceedings of the International Conference on Internet Computing in Science and Engineering. pp. 9-14.

[13] X. Yang, M. Zhang (2000) Necessary Constraints for Fusion Algorithms in Meta Search Engine Systems. In: Proc. of the Int'l Conf. on Intelligent Technologies. pp. 409-416.

[14] H. Jadidoleslamy(2012) Search Result Merging and Ranking Strategies in Meta-Search Engines: A Survey. IJCSI Int J Comp Sci 9(4):239-251.

[15] R. Albert, H. Jeong, A.L. Barabasi(1999) Diameter of the Worldwide Web. Nature 401:130-131.

[16] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. State, T.A. Wiener J (2000) Graph structure in the Web, Proceedings of the Ninth International Worldwide Web Conference, Elsevier.

[17] M.H. Marghny, A.F. Ali (2005) *Web Mining Based On Genetic Algorithm*. AIML 05 Conference, 19-21 December 2005, CICC, Cairo, Egypt.

[18] Z.Z Nick, P. Themis (2001) Web search using a genetic algorithm. Internet Comput 5(2):18-26. DOI: 10.1109/4236.914644.

[19] F. Picarougne, N. Monmarché, A. Oliver, G. Venturini (2002) Web mining with a genetic algorithm. Laboratoired'Informatique, Université de Tours,
64, Avenue Jean Portalis, 37200 Tours, France. Available at: http://www2002.org/CDROM/poster/58/. Accessed on: 2nd Jan 2014.

[20] A. Al-Dallal, R. Shaker (2009) Genetic Algorithm in Web Search using inverted index representation. GCC Conference and Exhibition, 17-19 March 2009, 5th IEEE. pp. 1-5. DOI: 10.1109/IEEEGCC.2009.5734301.

[21] H. Drias (2011) Web Information Retrieval Using Particle Swarm Optimization Based Approaches. wi-iat, vol. 1, pp.36-39. IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

[22] I.B. Priya,H.P Leena (2013) Web Information Retrieval Using Genetic Algorithm-Particle Swarm Optimization. Int J Future ComputCommun 2(6):595-599. DOI: 10.7763/IJFCC.2013.V2.234.

[23] R. Ghaemi, A.M. Fard, H. Tabatabaee, M. Sadeghizadeh (2008) Evolutionary Query Optimization for Heterogeneous Distributed Database Systems. World AcadSci Eng Technol 19:43-49. Available at: http://waset.org/publications/12865. Retrieved on 09th February 2014.

[24] Hjelmqvist, Sten (26 Mar 2012), Fast, memory efficient Levenshtein algorithm.Available at: (http://www.codeproject.com/Articles/13525/Fast-memory-efficient-Levenshtein-algorithm)

[25] S. García, J.R. Cano and F. Herrera (2008). A Memetic Algorithm for Evolutionary Prototype Selection: A Scaling Up Approach, *Pattern Recognition,* vol. 41, no. 8, pp. 2693-2709.

[26] H.A. Sanusi, A. Zubair , R.O. Oladele (2011). Comparative Assessment of Genetic and Memetic Algorithms. Journal of Emerging Trends in Computing and Information Sciences**.** VOL. 2, NO. 10, October 2011. ISSN 2079-8407

[27] P. Garg (2009). A comparison between memetic algorithm and genetic algorithm for the cryptanalysis of simplified data encryption standard algorithm. International Journal of Network Security & Its Applications, 1(1):34 -- 42, April 2009.

[28] A. Bookstein, D. Swanson (1976). Probabilistic models for automatic indexing.Journal for the American Society for Information Science 25(5):312 - 318,1976

[29] S. Wu, and J. Li (2004) Effectiveness Evaluation and Comparison of Web Search Engines and Meta-Search Engines. *WAIM, volume 3129 of Lecture Notes in Computer Science, page 303-314. Springer, (2004)*

[30] R. O Oladele (2013). Comparative Study Of Memetic Algorithm And Genetic Algorithm For Multi-Objective Optimization Of Network Design. International Journal of Emerging Trends & Technology in Computer Science. Volume 2, Issue 3, May – June 2013. ISSN 2278-6856.

L. Melita is the Assistant Professor and Chair-holder of the Information Systems Management Chair of the School of Computing, Jimma University, Jimma, Ethiopia. She is a scholar in Computer Science and has done her B.Sc in Madurai Kamaraj University, M.C.A in Manonmaniam Sundaranar University and her M.Phil in Madurai Kamaraj University. Her research interests are with Soft Computing, Bio-inspired Computing and Web mining.

She has more than 12 years of teaching experience and nearly 9 years of research experience. She has published 11 articles in national and international conferences and journals. She is a member of IEEE. She has received many prizes and awards from her schooling for her proficiency in academics.

Gopinath Ganapathy PhD is the Professor and Head of the School of Computer Science, Applications and Engineering, Bharathidasan University, India. He did his under graduation and post graduation in 1986 and 1988 respectively from Bharathidasan University, India. He obtained his PhD degree, in Computer Science in 1996, from Madurai Kamaraj University, India.He received the Young Scientist Fellow Award for the year 1994 and eventually did the research work at IIT Madras. He was a Consultant for 12 years in the international firms in the USA and the UK, including IBM, Lucent Technologies (Bell Labs) and Toyota. His research interests include Modeling, Patterns, NLP, Web Engineering, and Text Mining.

Gopinath has published more than 60 research papers. He is an editorial member for a few international journals. He is the "Fellow" of IACST, Senior Member in IEEE, ACM, Life Member in Indian Science Congress, ISTE and Computer Society of India and Chairman, Computer Society of India, Trichy Chapter. Currently he is the Director of Technology Park and the Chairman, School of Computer Science, Applications and Engineering, Bharathidasan University, Tamil Nadu.