Web Content Mining Using Evolutionary Algorithms: A Survey Report

Nafisha Mamti

Assistant Professor,
Department of Computer Engineering
Atmiya Institute of Technology and Science, Rajkot, India

***Abstract* - The web surfing has taken place in day to day work that leads to enormous mass of data over the web. The search engines helps to retrieve necessary data from massive databases over the internet. As each search engine has its own limitations to retrieve most relevant information that user is seeking for, user has to struggle to find interesting data from the results provided by traditional search engines. This problem is frequently faced when the user has given complex query i.e. many keywords in search box. To overcome this problem, the search results must be processed further in order to provide most relevant information in proper sequence. Many researchers addressed this issue and has found the various techniques among which the effective one is using evolutionary algorithms. This survey paper summerises such proposed solutions and enlightens how most relevant information can be produced for complex queries and remove noise. The evolutionary algorithms also used in web pages classification, clustering and feature selection.**

***Index Terms* - Web Content mining, Evolutionary Algorithms, Firefly Algorithm, Genetic Relation Algorithm, Memetic Algorithm, feature selection, web page classification, search engine**

I. INTRODUCTION

As per the increased use of internet, the data over web is rapidly increased in enormous number of databases. Usually user faces the problem in retrieving data that they are interested in. Even after getting results, they struggle to find what actually they want. Sometime, user may lack to give appropriate query or even after if user is a student, researcher or enough knowledgeable and give proper query then also they don't get accurate result for complex queries (i.e. more keywords). The search engines also provide bulk of search results that is impossible for users to go through each. As per the search pages increased, the irrelevancy of results to given query is also getting increased after certain pages [2]. The researchers introduces the solutions using evolutionary algorithms for search engines that help it to provide noise free information in proper priority to users. These solutions covers the classifications and clustering of web pages, best feature selection and processing over search results before producing it to users.

II. INTRODUCTION TO SEARCH ENGINES

The search engine is aimed to provide exact information to user as per given query which can be simple or complex. The search engine searches over indexed database for webpages which is filled by web crawler (also known as web spider). This database contains the terms of webpages and its links which are classified in order to increase searching speed. As per the analysis over search engines done by Prof. R. Arvindhan [1] traditional search engine has its own limitations which are I. poor interconnection of intended search information and the retrieved information. II Unable to ensure trust at all levels as it deals with enormous number of web users and web content. III. Lack of capability to understand the provided information. IV. Lack of Automatic information transfer.

III. WEB PAGE CLASSIFICATION USING MEMETIC ALGORITHM

In order to manage enormous data over web, the web pages are classified according their terms and quality factors measured by number of links it contain, amount of information, advertisements, hits etc. To give all web pages classification manually is tedious and impractical. It raises the need for automatic classification of web pages. The MA (Memetic Algorithm) based web documents classification algorithm is proposed by Xia Sun, Ziqiang WangDexian Zhang [3] which achieves better performance than other related classification algorithms.

Memetic algorithm combines evolutionary algorithms with the intensification power of a local search, and has a pragmatic perspective for better effects than GA. As such MA, a local optimizer is applied to each offspring before it is inserted into the population in order to make it towards optimum and then GA platform as a means to accomplish global exploration within a population. MAs are similar to GAs but the elements that form a chromosome are called memes, not genes. The unique aspect of the MAs algorithm is that all chromosomes and offsprings are allowed to gain some experience, through a local search, before being involved in the evolutionary process. From the algorithm procedure of memetic algorithm (MA), we can observe that the parameters involved in MAs are the same four parameters used in GAs: population size, number of generations, crossover rate, and mutation rate in addition to a local search mechanism.

The steps of the MA-based document classification algorithm are outlined as follows:

Step1: Documents feature selection is performed.

Step2: Individual encoding structure and Initialization. In the initialization process, a set of individuals (i.e.,chromosome) is created at random. The structure of an individual for document classification problem is composed of a set of term weight values. Therefore, an individual i can be represented as the vector $Xi = (x_{i1}, x_{i2}, \dots, x_{in})$ n is the number of term numbers in document sets.

Step3: Fitness function computation. As in all evolutionary computation techniques there must be some function or method to evaluate the goodness of an individual. In order to define a good fitness function that rewards the right kinds of individuals, we try to consider affecting factors as complete as possible to improve the results of classification. Two measures of effectiveness commonly used in text classification are adopted in this study, those are precision and recall. The former quantifies the percentage of documents that are correctly classified as positives (they belong to the category) and the latter quantifies the percentage of positive documents that are correctly classified. For evaluating the average performance across categories, the evaluation function F used in this study is based on averaging the F_1 value of each category in proportion to its number of documents, where F_1 is defined as follows:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (1)$$

$$Recall = \frac{Number\ of\ correct\ positive\ prediction}{Number\ of\ positive\ example} \qquad (2)$$

$$Precision = \frac{Number\ of\ correct\ positive\ prediction}{Number\ of\ positive\ predictions} \qquad (3)$$

```
Step4: Repeat.
For each individual $i \in P$: do local-search($i$);
For $i = 1$ to number of generations;
Select two parents at random $i_a$ and $i_b$;
Generate an offspring $i_c = Crossover\ (i_a$ and $i_b)$;
$i_c = local - search\ (i_c)$;
Calculate the fitness of the offspring in terms of Eq.(1);
If $i_c$ is better than the worst chromosome
Then replace the worst chromosome by $i_c$;
Next i;
Step5: Stop condition. Check if the iteration number approaches to the predefined maximum iteration;
```

IV. Feature selection using firefly algorithm

As per mentioned in given algorithm of web page classification, the first step is to find the subset of features in web pages in order to classify them in most relevant category. According to research done by Esra Saraç and Selma Ayşe Özel, [4] the Firefly Algorithm can be used to find best feature from given web page or web document and also a set of best features can be produced.

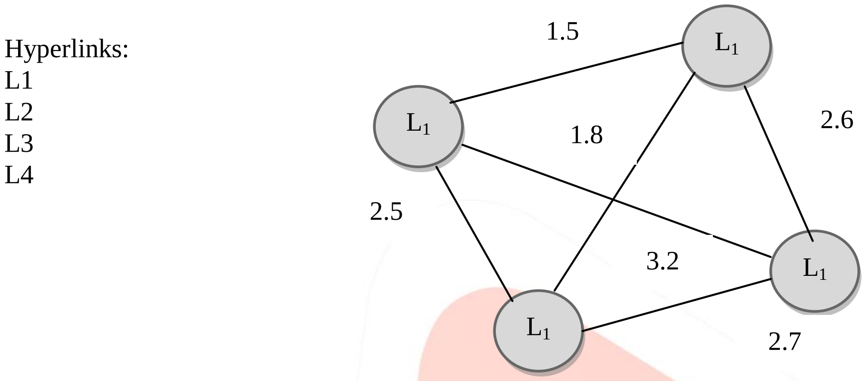
Firefly Algorithm (FA) is an optimization technique, developed recently by Xin-She Yang at Cambridge University. It is inspired by social behavior of fireflies and the phenomenon of bioluminescent communication. Fireflies can generate light inside of it. Light production in fireflies is due to a type of chemical reaction. The pattern of flashes is unique for a particular species of fireflies. However, two fundamental functions of such flashes are (i) to attract mating partners or communication, and (ii) to attract potential victim. Flashing may also be used for a protective warning mechanism. The light intensity at a particular distance from the light source follows the inverse square law. It means that, as the distance increases, the light intensity decreases. Furthermore, the air absorbs light which becomes weaker and weaker as there is an increase of the distance. There are two combined factors that make most fireflies visible only to a limited distance that is usually good enough for fireflies to communicate each other. The flashing light can be formulated in such a way that it is associated with the objective function to be optimized. This makes it possible to formulate new metaheuristic algorithms. The main steps of FA described below:

```
1.  Generate initial population of fireflies ($x_i$)
2.  Determine light intensity $I_i$ for each firefly by their $df(i) = df(i) * exp^{-\lambda * F-measure}$ values
3.  While (t < MaxGeneration) do
    3.1.  for each firefly $x_i$ do
                for each firefly $x_j$ do
                    if ( $I_j > I_i$ ) then
                        Move firefly i towards j in d-dimension
                    endif
                Update attractiveness with respect to F-measure
                Evaluate new solution and light intensity
              end for
          end for
    3.2.  Rank the fireflies and find the current best one
    3.3.  Increment t
    end while
4.  Display the best firefly
```

V. NOISE REMOVAL USING GENETIC RELATION ALGORITHM

The search engine produces bulk of data where the massive amount of data are irrelevant as per increased number of search pages that raises the need of additional filter to searched results. The most relevant hyperlinks can be found using genetic relation algorithm (GRA) that is modified algorithm of genetic algorithm by Eloy Gonzales [2] as described below:

Genetic Relation Algorithm is one of the evolutionary optimization techniques, which evolves both the directed and indirected graphs, where the best relations between events are obtained. Therefore, it is used to extract a fairly small number of events from a large set of events. An event is an abstract concept being represented and encoded by nodes in GRA. The events are defined depending on the problem to solve, for instance, in the stock market, the nodes represent stock brands in a portfolio, in data mining, the nodes represent association rules and in web mining, the nodes represent hyperlinks of a web page. The strength is defined between the nodes and it is used in the fitness function. The cosine similarity between nodes is used as the strength. The following figure shows the example of GRA.



Node function hyperlinks (i,j) of the webpage
Strength: Similarity between nodes

Fig. 1 Genetic Relation Algorithm for Web Mining

Table 1 genotype of GRA for web mining

NodeNo	NodeFunction	. Ci1	Ci2	Ci3	Si1	Si2	Si3
1	L1	2	3	4	1.5	1.8	2.5
2	L2	1	3	4	1.5	2.6	3.2
3	L3	1	2	4	1.8	2.6	2.7
4	L4	1	2	3	2.5	3.2	2.7

The query (q) can be words, word pairs or phrases given by users and it is defined as a list of keywords (or terms). L is the set of suffixes of input keywords in query. Node i is represented as a term vector as follows:
Node $i = (d_{i1}, d_{i2}, \dots, d_{il}, \dots, d_{i|L|})$
Where d_{il} represents the occurrence of term l in node i.

$$d_{il} = \begin{cases} 1 \; if \; node \; i \; contains \; term \; l \\ \quad 0 \; otherwise \end{cases} \qquad (4)$$

The quality of node i is calculated as follows:

$$F(i) = \sum_{l \,\epsilon\, L} d_{il} \qquad (5)$$

The cosine similarity between node i and node j is calculated as follows:

$$D(i, j) = \frac{\sum_{l \in L} f_{il} f_{jl}}{\sum_{l \in L} f_{il} \sum_{l \in L} f_{jl}} \qquad (6)$$

where, f_{il} is the frequency of keyword l in node i, Cosine measure determines the similarity between two vectors (nodes) independently of their magnitude. Eq. 6 returns the angle between these two vectors. It is equal to 1 when the vectors point in the same direction, and zero when they form a 90 degrees angle.

Using GRA, an efficient searching strategy can be defined. That is, GRA is used to find the most interesting pages for users.

Since the number of nodes in GRA defines the number of hyperlinks, GRA selects a reduced number of hyperlinks which best match the query generated by users.

A. Fitness of GRA

The fitness function of GRA individuals is defined as follows:

$$Fitness = \frac{1}{|R|} \sum_{i \in R} \frac{1}{|R(i)|} \sum_{i \in R} F(i,j) \qquad (7)$$

$$F(i,j) = D(i,j) * F(i) * F\,(j) \qquad (8)$$

where,
D(i, j): cosine similarity between node i and node j described in the previous section.
F(i): quality of node i.
R: set of suffixes of nodes (hyperlinks) in GRA.
R(i): set of suffixes of nodes whose similarity is defined between node i in GRA.

The fitness function evaluates the GRA individuals so that the similarities between hyperlinks are maximized. The proposed method starts with a query request generated by users. Then, using standard searching engines (ex. Google ,etc.), the initial population of GRA individuals is created randomly from the results obtained by the search engines. That is, GRA nodes encode the hyperlinks generated by any standard search engine. Then, the GRA individuals are evaluated with the fitness function and genetic operations (elite selection, tournament selection, crossover and mutation) are carried out. Finally, the final elite individual is the output given to the user.

B. Evaluation Measures

The contents of the function nodes of the final elite individual are listed from the first to the last node to form a rank of hyperlinks.

The evaluation of the final elite individual is realized using the conventional evaluation measures, i.e. precision, recall and F-score for all nodes as follows:

Let N be the set of hyperlinks retrieved by any standard search engine. Let Dq be the set of actual relevant hyperlinks of query q in N.

Recall of node i is defined as:

$$r(i) = \frac{s_i}{|D_q|} \qquad (9)$$

where si is the number of relevant hyperlinks from node 1 to node i. si ⩽ |Dq|

Precision of node i is defined as:

$$p(i) = \frac{s_i}{pos(i)} \qquad (10)$$

Where pos (i) is the position of node i, referred as ranked node number.

Average precision is computed based on the precision of each node which contains a relevant document as follows:

$$p_{avg} = \frac{\sum_{i \in D_q} p(i)}{|D_q|} \qquad (11)$$

Fscore of node i is defined as follows:

$$FS(i) = \frac{2p(i)r(i)}{p(i)+r(i)} \qquad (12)$$

According to experiments done using above algorithm the GRA outperforms the conventional web search engine especially between the recall levels of 40% and 50%. The preliminary results demonstrated that the system is capable of selecting the most interesting hyperlinks from the web related to a query. It is important to remark that the usefulness of an important hyperlink is related to all relevant words contained in the query.

VI. Conclusion

This paper address various evolutionary algorithm to increase the quality of search results. It is complementary to search engine as it doesn’t replace the search engine whole, instead it supports by various techniques. Collectively these algorithms will give best output rather than individual. User can retrieve most interesting information even for any complex queries in limited interested results that means removal of unnecessary information.

REFERENCES

[1] R Aravindhan, and Dr.R.Shanmugalakshmi, “Comparative Analysis of Web 3.0 Search Engines: A Survey Report,” 2014 International Conference on Computer Communication and Informatics (ICCCI -2014), Jan. 03 – 05, 2014, Coimbatore, INDIA, 9781479923526/14/$3100 ©2014 IEEE

[2] Eloy Gonzales, Shingo Mabu, Karla Taboada and Kotaro Hirasawa, “Web Mining using Genetic Relation Algorithm”, SICE Annual Conference 2010 August 18-21, 2010, The Grand Hotel, Taipei, Taiwan, ¥400 © 2010 SICE PR0001/10/00001622

[3] Xia Sun, Ziqiang Wang, Dexian Zhang, “A MA-Based Web Document Classification Algorithm”, Proceedings of 2008 IEEE International Smposim on IT in Medicine and Edcation, 9781424425112/08/$2500 ©2008 IEEE

[4] Esra Saraç, Selma Ayşe Özel, “Web Page Classification Using Firefly Optimization”, 9781479906611/13/$3100 ©2013 IEEE