*Journal of **AI** and **D**ata **M**ining*
Published online:

A technique for improving web mining using enhanced genetic algorithm

Fatemeh Nosratian, Hossein Nematzadeh*, Homayun Motameni

Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran

Receive Date; Acceptance Date

Corresponding author: nematzadeh@iausari.ac.ir (H.Nematzadeh)

Abstract

World Wide Web is growing at a very fast pace and makes a lot of information available to the public. Search engines used in the conventional methods to retrieve information on the web; however, the results of these engines are still capable of being refined and their accuracy is not high enough. One of the methods available for web mining is evolutionary algorithms, which do searches according to the users' interests. A new genetic algorithm (GA) optimizes the important relationships among links on web pages with evolutionary algorithms. This paper presents a new method for classifying web documents based on the modified GA to find the best pages among the ones searched by engines. It also calculates, independently or dependently, the quality of pages by web page features. The proposed algorithm is complementary to the search engines. In the proposed method, after implementation of GA using the MATLAB 2013 software and a cross-over rate of 0.7 and a mutation rate of 0.05, the best and most similar pages are presented to the user. In the case of algorithm recurring, the result will not change.

Keywords: *Genetic algorithm, web miming, evolutionary computation.*

1. Introduction

Web dramatic expansion in a decentralized and disorganized process has led to the establishment of a huge amount of information and documentation related to each other, and has brought a great deal of challenge for its users. In fact, the web is composed of a large complex set of structured and unstructured data. In order to improve the search results, the genetic algorithm (GA) techniques are used [17]. In addition, the weighting criteria and the use of algorithms navigating the users' information are useful techniques to offer users the best pages. Technical GA is based upon natural selection and genotypes. In this paper, the results of the search engines is formed as a large chromosome (a string of bits, each representing a web page), and then broken into smaller chromosomes (strings smaller bits) in order to reduce their computational size, and GA is applied to small units to obtain a reduced reasonable computational volume. According to the properties of the dataset, the genes attributed to chromosomes are determined, and evolutionary computations are applied to them. This paper is based upon a dataset with four features used as genes. Also the evaluation of web pages is based upon these features. This paper considers a method for web mining to be improved using GA and to offer pages with a higher quality to the users. To increase the functionality and speed in computing GA, we decided to improve it and make changes in it. This paper is a way for web mining plans. We also tried to introduce a better fitness and evaluation functions through normalizing genes in both the local and global levels in order to increase the accuracy of the final response.

This paper is organized in four sections. In Section 2, the related studies are presented. In Section 3, the proposed algorithm is proposed. Section 4 provides relevant conclusion.

2. Related works

Twycross and Cayzer [10], in their paper, have presented a system that learns the users' interests with a set of web pages ranked by the user. They use this

system to determine if other web pages were irrelevant and relevant to the user. The method of the study is that the system learns irrelevant or relevant concepts using a list of pages visited by the user and whether or not the pages are associated with his work. After learning, it is given these concepts. Then based upon these pages ranked by the user, the system uses those for ranking pages not viewed by the user yet. Therefore, it helps the user searching. In order to build such a system, Twycross and Cayzer used a collaborative development of an evolutionary algorithm network for optimization of neural functions and used consecutive decision rules and learning algorithm to develop some sub-types. These sub-types, which are internally proliferated, are formed by some units. Each unit only shows a part of the solution, and they are combined to obtain a final solution. This classification shows the best combination pages. In order to validate their proposed system, various standard methods such as Naive Bayes, nearest neighbor, decision tree, and neural networks were compared. The simple Naive Bayes method shows the best results. In this regard, they compared their implementation results just with the Bayesian method.

In [2], a simple model has been provided for clustering web users. In this model, the users' interest to a web page function is estimated using the user elapsed time on that page. In this way, a lot of useless data is removed from the sample space. The proposed method identifies noteworthy web pages which have been specified by user. These pages constitute the search history of users. New web pages are then suggested based on user history.

An evaluation of this approach has shown that in comparison to the existing search engines, the satisfaction of the studied users has increase regarding the compliance of the test results ranking with their interests.

In [12], a method has been tried to personalize solutions through re-rank by adding a new variable to the personalization issue of the peer-to-peer information retrieval system. This method increases the system scalability using the data recovery algorithms that use the cooperation methods. This method is exclusive and flat. In 2004, various forms have been mentioned for this algorithm but all of them are repetitive and try to estimate the following cases for a fixed number of clusters:

(A): Obtaining points as cluster centers; these spots in fact are the average points of each cluster.

(B): Assigning each data sample to a cluster that has the shortest distance to the center of the cluster. In a simple form of this method, some points are randomly selected based on the number of clusters required. Then the data can be assigned to one of these clusters regarding their proximity (similarity), and the new clusters are achieved. By repeating this procedure, new centers can be calculated for them in each repetition by averaging the data, and the data is re-attributed to a new cluster. This process continues until a change in the data is not reached. In all of them, taking into account the web structural information, the users' survey information take place.

Yang and Chen have presented a method for modeling the structure of the web using Petri net in addition to introducing Petri net as a high-level graph used in modeling the activities of simultaneous systems [7]. In this method, the locations represent web pages on the site and the transitions are representative of links between pages. This paper focuses on how to use the GSM algorithm for retrieving the contents of web pages, content analysis, and finding a matrix that represents the structure of the web. It also shows how to identify the main page and to complete the process with the availability feature. Using the Markov analysis, the statistical information of using pages for discovering patterns is also considered.

Chen et al. have analyzed the web after introducing random timed Petri nets [11]. In order to facilitate the data pre-processing phase and to improve the accuracy of the results obtained in the web mining process, the web structure was modeled. In some articles, personalization and the need for it have been studied. They explain that this is a website selection according to the needs of specific users, and refers to the type and information display on the web, and is provided according to the history stored from the web application. Web users with different interests and tastes are colloquially using it.

Bautista et al. [13] have suggested a method with genetic algorithm that processes retrieval information by genetic fuzzy classification and genetic feature selection, and evaluates documents for a user based on keywords. Two main models in this system are genetic feature selection and fuzzy classification. The problem complexity is eliminated by unrelated features due to the feature selection. This method increases the quality of the query.

Hossaini et al. [14] have used genetic algorithms to classify and cluster, and also have worked on variable size vectors. They combined mutation and intersection standards in the genetic algorithm and divided the result by K-means algorithm and improved it. In this method, there are some classes and sub-classes. In this method, the accuracy rate increases. Eloy Gonzalez [15] has used GRA for web mining. The undirected graph was used in this method and the connections between pages on a web were investigated. In this method, there also exist intersection and mutation and connections, and the

contents of each node is randomly mutated with Pm rate and blended with Pc rate. The quality of nodes is calculated based on their connections. The cosine similarity function is used to calculate the similarity between two nodes. The fitness function is defined as follows:

$$Precision = (\#relevant \cap retrieved)/(\#retrieved) \quad (1)$$

$$Recall = (\#relevant \cap retrieved)/(\#relevant) \quad (2)$$

The evaluation criteria in this method are accuracy, overlap, and F-score. More et al. [17] have used evolutionary algorithms for web mining. In this method, the result of the inquiry from google API was collected and then processed using the memetic algorithm. Using the local search and discovery function, the result of the inquiry is added to the nearest cluster. The relationship between the results is measured using cosine similarity, and memetic algorithms are used for optimization. Heuristic function in this method measures snippets quality and also determines a value for repeat threshold for the total effective weights on snippets in exploratory functions. Finally, the two criteria of overlap accuracy are calculated.

3. Proposed Method

This paper is aimed to get pages with the best quality using altered genetic algorithm. The best quality is for pages with the most similarity to the subject of search. Thus in this way, the resulting proposed pages from genetic algorithm are most similar to the subject of the search. This project presents a new method to classify chromosomes of the genetic algorithm via matrix modification, and proposes functions for assessing the quality of gene functions and the degree of pages' similarity. Finally, a method is offered to gain fitness function for chromosome quality. It should be noted that this method is not an alternative to web searches but is complementary for the best pages to be offered to the user. In the genetic algorithm, the input is the initial search done by the search engine and in fact is the producer of the initial population to be used in the genetic algorithm. In this way, if we want to propose n best pages out of m existing pages with common genetic algorithm, a $m \times n$ matrix should be considered, where m refers to the number of pages found by the search engine and n refers to the beat pages regarding their quality and similarity. One of the big problems of such calculation method is that it is long-lasting and not affordable considering time and cost because in this method, the chromosome size in the genetic algorithm is equal to m. For example, if m (the number of pages searched by the search engine) is equal to 1,000,000 and n (number of pages with superior similarity and quality) is considered to be 1000, then the matrix $m \times n$ is equal to 1,000,000 × 1000, and this matrix has 10^8 members and the size of each chromosome is 1,000,000, and carrying out cross-over and mutation and calculating the quality of the pages and applying, the evaluation functions

$$1000\text{chromosems}\begin{cases} c_1, \dots, c_{10000}, \dots, c_{1000000}, \dots, c_{1000000} \\ \vdots \\ c_1, \dots, c_{10000}, \dots, c_{1000000}, \dots, c_{1000000} \end{cases}$$

Figure 1.The structure of chromosomes in GA

on this matrix are very time-consuming and costly (Figure 1). In the example above, if the number of chromosomes is 1000 and if each chromosome has 1000000 genes, computing such long chromosomes requires a lot of time and high cost because while doing genetic algorithm in such chromosomes with a high number of genes after cross-over, the quality of new chromosomes should be re-assessed. In this case, if each gene has four chromosomes, then the number of calculations run for the chromosome quality is:

$1{,}000{,}000 \times 4 = 4{,}000{,}000$

$4{,}000{,}000 \times 1000 = 4{,}000{,}000{,}000$

This is the number of estimates of the fitness function for chromosomes. In the previous methods, the two functions accuracy and overlapping are used for evaluating the chromosomes. They are two measurement criteria only with regard to the keywords.

3.1. Proposal

In this work, a large chromosome is divided into small parts to eliminate the massive matrix with big elements so that each part is a new chromosome (Fig. 2). The number of divisions is based on the number of pages proposed to the user. The matrix in Figure 3 is obtained after such segmentation.

$$m\begin{cases} \{m_1, m_2, m_3, \dots, m_{n-1}, m_n\} \\ \{m_{n+1}, m_{n+2}, m_{n+3}, \dots, m_{2n-1}, m_{2n}\} \\ \vdots \\ \{m_{kn+1}, m_{kn+2}, m_{kn+3}, \dots, m_{kn-1}, m_{kn}\} \end{cases}$$

Figure 2. Position of genes in chromosome with the proposed method

$$\begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{\frac{m}{n},1} & \cdots & a_{\frac{m}{n},n} \end{bmatrix}$$

Figure 3. Exposure matrix chromosomes

where m is the initial chromosome size.

In this case, a chromosome is broken and divided into smaller chromosomes. The number of divisions depends on the chromosome initial size and the number of pages proposed to the user. The division method is that the initial search pages are formed as a chromosome. Due to the length of the chromosome, it is divided into smaller numbers so that the size of the new chromosomes is equal to the number of superior pages proposed to the user. For example, if the initial searched pages are 2000, i.e. the initial chromosome length is 2,000, and if it is supposed that 10 top pages, in terms of quality, are to be

proposed to the user, they are broken into 200 chromosomes with the length of 10.
Each chromosome has n genes and each gene used in this work considering the population has four features that are common to the population of the searched pages (Figure 4).

- The first feature is the number of lines in each page. The more the number of lines is associated with the keyword, the greater the quality will be.
- The second feature is the group that shows the belongings of one page to different groups.
- The third feature is the number of referrals to the page. The more referrals to a page indicate a higher quality and importance of the page.
- Finally, the fourth feature is the category or organization to which the page has been attributed.

Gene1				...	Gene n			
Line	group	ref	Organize	...	Line	group	ref	Organize

Figure 4. N genes with four features for chromosome

In order to obtain a valid and normalized value for each chromosome, we used the following method:
The quality of pages and the degree of similarity between the pages of each chromosome depend on the total quality of genes of that chromosome. Each chromosome has n genes and each gene has four features (Fig. 5). The quality of genes depends on the total ratio of their features in the same chromosome. The quality of each gene feature is separately calculated as follows:

$$\text{chromosome } i \begin{cases} \{line_1, line_2, \ldots, line_n\} \\ \{group_1, group_2, \ldots, group_n\} \\ \{ref_1, ref_2, \ldots, ref_n\} \\ \{organiz_1, organiz_2, \ldots, organiz_n\} \end{cases}$$

3.2. Evaluation Function

The evaluation function is used to assess the quality of the parent genes before applying the genetic algorithm, and the quality of the child genes is also analyzed, after applying cross-over and mutation, to determine if the algorithm application improves the population or not. For each group of gene features, the evaluation function is accordingly used. Thus:

$$l = \sum_{i=1}^{n} line_i \quad (3)$$

$$fl_i = \frac{line_i}{l} \quad (4)$$

fl_i: Quality ratio to the number of keywords

$$grp_i = \frac{\sum group_i}{\sum group_1, group_2, \ldots, group_k} \quad (5)$$

$$f_{grp_i} = \frac{\sum group_x + \sum group_y + \sum group_z}{\sum group_1, group_2, \ldots, group_k} \quad (6)$$

f_grp_i: Quality of dependency similarity of one page to another on the same chromosome proportional to the category
$\sum org_i$: Total number of groups to which one page belongs, and is similar to other pages on the same chromosome. In fact, to determine the groups' similarity rate in a chromosome, the comparison should be done locally.

$$ref = \sum_{i=1}^{n} ref_i \quad (7)$$

$$f\_ref_i = \frac{ref_i}{ref} \quad (8)$$

f_ref_i: Quality of a page based on the number of visits

$$forg_i = \frac{\sum org_i}{\sum org_1, org_2, \ldots, org_k} \quad (9)$$

$forg_i$: It shows the similarity dependency of one page to one category with other pages on the same chromosome.

$$f\_gen_i = f\_l_i + f\_grp_i + f\_ref_i + f\_org_i \quad (10)$$

f_gen_i: Quality of one gene

For example, see Table 1. In this case, if the table is part of a 2000-times search and we want to consider the top ten pages in terms of quality and similarity, it includes two chromosomes that become the number of page genes, and each gene has four features: the number of lines, group name, frequency of referrals, and organization name (Figure 6).

Gene 1				.. .	Gene10			
1948	$g_1g_2g_3$	122	org 1	.. .	421	$g_3g_{19g_5}$	142	org 6

Figure 5. View of a chromosome with a feature gene

Gene 11				.. .	Gene 20			
969	$g_2g_{16}g_3$	217	org 9	.. .	1015	$g_3g_{19}g_2$	597	org 3

Figure 6. Chromosome with ten genes

$$l = \sum_{i=1}^{n} line_i = (1948 + 1991 + \ldots + 421) = 20655$$

$$fl_1 = \frac{line_1}{l} = \frac{1948}{20665} = 0.094$$

$$fl_2 = \frac{line_2}{l} = \frac{1991}{20665} = 0.096$$

fl_2: It is a greater number compared to fl_1, and indicates that the gene Page 2 in terms of the number of lines associated with the keyword is of a higher quality. Normalization in this case is global.

$$grp_{g_1} = \frac{\sum group_1}{\sum group_1, group_2, \ldots, group_k} = \frac{\sum g_1}{\sum g_1, g_2, g_3, g_4, g_5, \ldots} = \frac{3}{30}$$

$$grp_{g_2} = \frac{\sum group_1}{\sum group_1, group_2, \ldots, group_k} = \frac{\sum g_2}{\sum g_1, g_2, g_3, g_4, g_5, \ldots} = \frac{1}{30}$$

$$f\_grp_1 = \frac{3}{30} + \frac{1}{30} + \frac{4}{30} = \frac{8}{30}$$

$$f\_grp_2 = \frac{3}{30} + \frac{2}{30} + \frac{4}{30} = \frac{9}{30}$$

f_grp_2: It is greater than f_grp_1, and indicates the higher quality of gene 2 compared to the groups belonging to it, i.e. pages of gene 2 have more similar groups than gene 1. Normalization in this case is

local, which represent a higher quality of gene 2 than gene 1.

$$ref = \sum_{i=1}^{n} ref_i = 1223 + 287 + 1050 + \ldots + 597 = 18062$$

$$f\_ref_1 = \frac{ref_1}{ref} = \frac{1223}{18062} = 0.0677$$

$$f\_ref_2 = \frac{ref_2}{ref} = \frac{287}{18062} = 0.015$$

$$forg_1 = \frac{\sum org_1}{\sum org_1, org_2, \ldots, org_k} = \frac{1}{10}$$

$$forg_2 = \frac{\sum org_2}{\sum org_1, org_2, \ldots, org_k} = \frac{3}{10}$$

$$f_{gen_1} = fl_1 + fgrp_1 + ref_1 + forg_1 = 0.094 + 0.26 + 0.067 + 0.1 = 0.521$$

$$f_{gen_2} = fl_2 + fgrp_2 + ref_2 + forg_2 = 0.096 + 0.3 + 0.015 + 0.3 = 0.711$$

$f\_gen_2 = 0.711, f\_gen_1 = 0.521$ Indicative of a higher quality than the first gene is the second gene

3.3. Fitness Function

The Fitness function is the evaluation of the goodness of chromosomes in the genetic algorithm, and selects a weight for each page on the basis of the keyword, number of referrals, similarity of genes in a chromosome, and assigned groups and categorizing them. If the page is a parent page having a higher weight compared to the child page, it remains in the original population; otherwise, the child with a higher quality will be placed in the population instead of the parent (Figure 7).

$$\begin{cases} parent1 \\ parent2 \end{cases} \xrightarrow{crossover\ ,\ mutation} \begin{cases} parent1 \\ parent2 \\ child1 \\ child2 \end{cases} \xrightarrow{select} \begin{cases} parentnew1 \\ parentnew2 \end{cases}$$

Figure 7. Method for selecting superior chromosomes

$$f_{chromosomej} = \sum_{i=1}^{n} f\_gen_i \quad (11)$$

$$f_{cromosome1} = f_{gen_1} + f_{gen_2} + \ldots + f_{gen_{10}} = 0.68 + 0.81 + \ldots + 0.655$$

In this work, after selecting two chromosomes for cross-over and mutation and creating two new children from parents, their genes were analyzed and

Table 1. 20 genes with their features

page number	Number of line	Group name			Number of visits	Organization name
1	1948	g_1	g_2	g_3	1223	org1
2	1991	g_4	g_5	g_3	287	org9
3	309	g_6	g_7	g_1	1050	org9
4	1539	g_8	g_9	g_{10}	1242	org3
5	1102	g_4	g_6	g_{11}	828	org7
6	433	g_{12}	g_{13}	g_{14}	1245	org5
7	1158	g_1	g_{15}	g_4	157	org9
8	741	g_{16}	g_{16}	g_{17}	582	org6
9	1994	g_8	g_{18}	g_3	1562	org3
10	421	g_3	g_{19}	g_5	1427	org6
11	969	g_2	g_{16}	g_{13}	217	org9
12	1008	g_{18}	g_{20}	g_2	1654	org7
13	1386	g_1	g_{17}	g_5	873	org8
14	595	g_1	g_4	g_{13}	1694	org4
15	1016	g_{10}	g_{12}	g_1	709	org5
16	958	g_{11}	g_1	g_4	460	org9
17	1354	g_{13}	g_{18}	g_6	1077	org3
18	1461	g_4	g_8	g_{20}	1017	org7
19	1501	g_{16}	g_3	g_4	1588	org4
20	1015	g_3	g_{19}	g_{12}	597	org3

PARENT CHROMOSOM 1	1	2	3	4	5	6	7	8	9	10
PARENT CHROMOSOM 2	11	12	13	14	15	16	17	18	19	20

Figure 8. Two chromosomes selected with ten genes

CHILD CHROMOSOM 1	1	2	3	4	5	6	17	18	19	20
CHILD CHROMOSOM 2	11	12	13	14	15	16	7	8	9	10

Figure 9. Cross-over

CHILD CHROMOSOM 1	1	2	3	4	5	6	17	18	19	20
CHILD CHROMOSOM 2	11	12	13	14	15	16	7	8	3	10

Figure 10. Mutation

CHILD CHROMOSOM 1	1	2	3	4	5	6	7	8	9	10
QUALITY GENES	.631	.702	.629	.66	.493	.401	.684	.488	.629	.655
CHILD CHROMOSOM 2	11	12	13	14	15	16	17	18	19	20
QUALITY GENES	.44	.656	.544	.737	.495	.631	.595	.665	.759	.516
CHILD CHROMOSOM 1	1	2	3	4	5	6	17	18	19	20
QUALITY GENES	.631	.702	.629	.66	.493	.401	.595	.665	.759	.516
CHILD CHROMOSOM 2	11	12	13	14	15	16	7	8	9	10
QUALITY GENES	.44	.656	.544	.737	.495	.631	.684	.488	.629	.655

Figure 11. The quality of parent and child chromosomes

evaluated in accordance with the evaluation function and the fitness function. Then the two top chromosomes were chosen after determining the weight of each page. For example, we considered the two chromosomes 1 and 2 in Figure 8.

After cross-over and random mutation, the new children are produced (Figs. 9 and 10).

The quality of the parent and child chromosomes has been calculated (Figure 11) and the two top chromosomes between parents and children are selected based on Equation (11).

$$f_{\_chromosome1} = \sum_{i=1}^{n} f_{gen_i} = 6.227$$

$$f_{\_chromosome2} = 6.058$$

In this example, the parents' quality is higher than that for the children, and no substitute occurs for the new children. After running cross-over with a rate of .7 and a mutation rate of 0.7, the remaining population in terms of quality and similarity is in the highest position so that with the selection of any chromosome of all the existing chromosomes, the best pages in terms of quality and similarity are recommended to the user. The selection of each chromosome has no effect on the outcome because the quality of chromosomes remains the same after several generations at a determined rate. The general method is shown in Figure 12.

1. The initial population is that of the search results.
2. Division is a proposed method presented in Equations (1) and (2).
3. Two chromosomes are selected using the roulette wheel.
4. The cross-over is applied on the selective chromosomes at a rate of 0.7. Each chromosome gene is randomly selected and the cross-over is a single point.
5. At this point, the mutation is run with a rate of 0.05 on genes randomly selected.
6. After performing the above steps, the quality of new pages is determined. In the case that the quality of new children is higher than their parents, their parents will be replaced.
7. The perquisite of the cycle completion is applying all the genetic algorithm rules and running cross-over and mutation with a determined rate.
8. At the end, some chromosomes of higher quality will remain in the matrix.
9. In fact, they are the pages proposed to the user.

3.4. Results and Discussions

In this work, the MATLAB 2013 software was used to analyze the data in the studied population. MATLAB is a software for data processing with a high-level language. In fact, the MATLAB software is the matrix laboratory, in which even individual numbers are considered as a matrix. In fact, all the data is stored in MATLAB in the form of a matrix. The MATLAB software has unique features not present in other programming languages. Among the benefits of this software are quick and easy coding with a high-level language, simple problem solving, simple user environment, allowing easy 2D and 3D figures and graphical representation of the results.

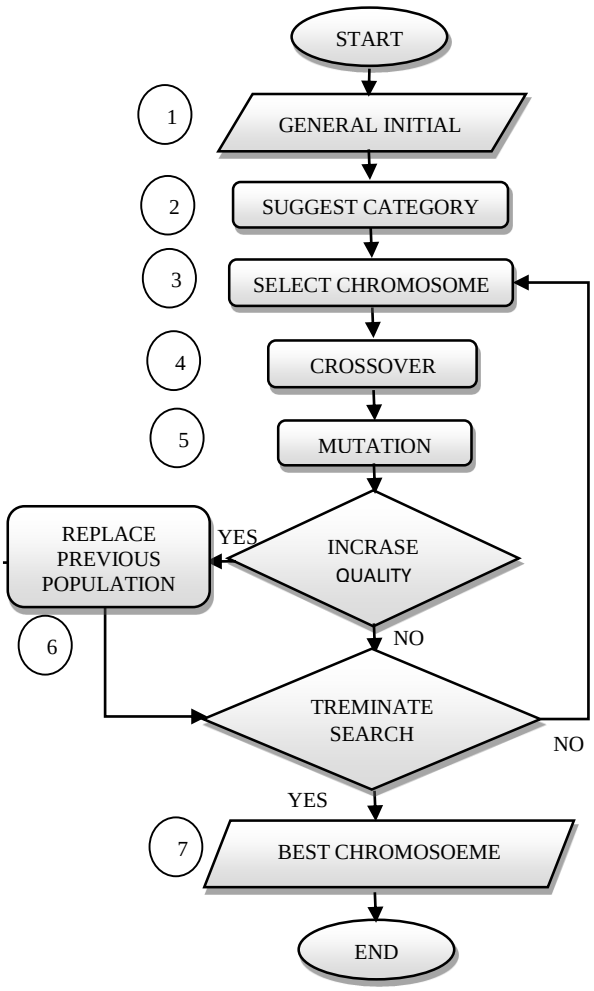

Figure 12. Flowchart of the proposed method

The following results were presented after coding the proposed algorithm in MATLAB codes and running the proposed method and defined functions in this language:

The program was first set to 50-times repetition of the genetic algorithm and tested with 2000 data (www.data.news20.tar.gz) collected in 1998. In each run of genetic algorithm, the maximum time for calculation was one minute. At the end, the 10 top pages of 2000 should be selected. After 50 generations, Figure 13 was obtained. The horizontal axis shows the number of generations, and the vertical axis shows the quality of the pages. The more the number of generations are, the greater the page qualities will be but this still is not stable, and we finally produced heterogeneous chromosomes in terms of quality and similarity.

We tested the program with 100 repetitions, and 2000 data and chromosomes must have the best pages in terms of quality and similarity at the end. After 100 times cross-over and mutation, Figure 14 was obtained. In this case, the quality of pages increased compared to 50 times. At the end of

generation, an approximate stability was achieved but there were still differences among the selected chromosomes.

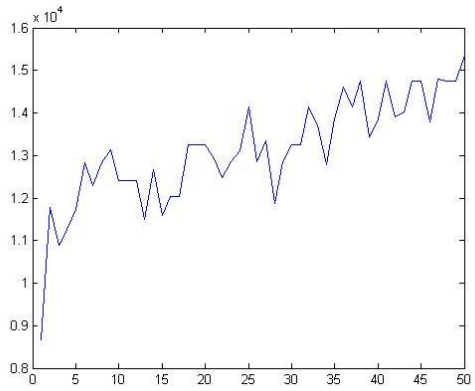

Figure 13. 50 generations

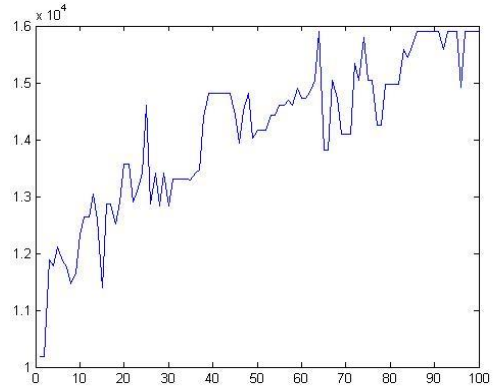

Figure 14. 100 generations

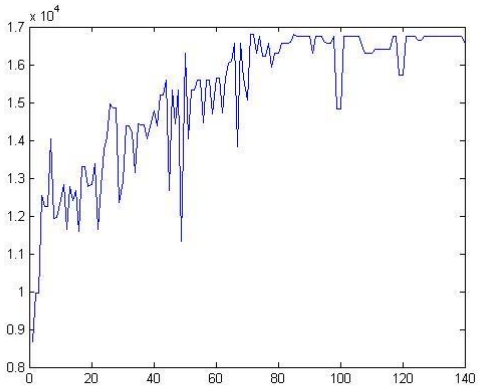

Figure 15. 140 generations

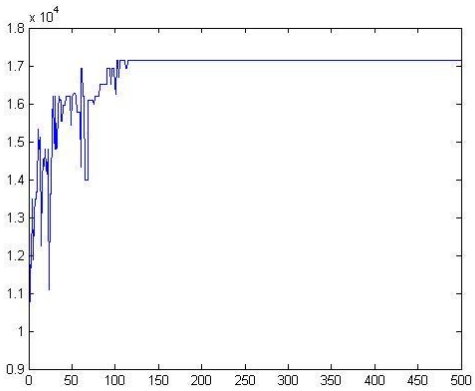

Figure 16. 500 generations

The program was tested after 140 times using 2000 data. At the end of the program, the chromosomes are more similar to each other, and all have the best page of similar requirements (Figure 15). In this case, the quality of pages increased compared to 50 and 100 times. At the last generations, fixed genes have remained in the population. This shows that chromosomes of higher quality and similarities have remained at last.

This section is to compare the results of the proposed algorithm with other algorithms. The proposed algorithm is compared with three other algorithms. This comparison is performed by the three measures precision, recall, and F_score, which are as what follow:

Precision of node I is defined as:

$$precision_i = \frac{f\_gen_i}{position_i}$$

f_gen_i: Quality of node i.$Position_i$ is the position of node I that is referred to as the ranked node number.

$$Recall_i = \frac{f\_ref_i}{\sum_{j=1}^{n} f\_ref_j}$$

$fref_i$: Number of references to node i.

F-score of node i is defined as follows:

$$F\_score = \frac{2p_{(i)} \times r_{(i)}}{p_{(i)} \times r_{(i)}}$$

The F_score function evaluates the selected pages, and illustrates the proposed method. The greater recalls yield more retrieval. Also a higher quality yields a higher precision. The F_score function is to compromise between recall and precision.

The following graphs show the F_score function for the proposed algorithm and the others.

Figures 17-19 are graphs comparing the average f_score of the proposed algorithm with three other algorithms. Figure 17 shows that the proposed algorithm is higher than the intelligent web mining technique using the evolutionary algorithms [17]. Table 2 compares the average F_score in different algorithms.Figure16 is obtained after 500 times, and represents the constant quality of pages after 140 generations. As a result, the continuation of the proposed algorithm is not useful after 140 generations. According to a cross-over rate of 0.7 and a mutation rate of 0.05, 140 times of generation is enough and proper for 2000 data.

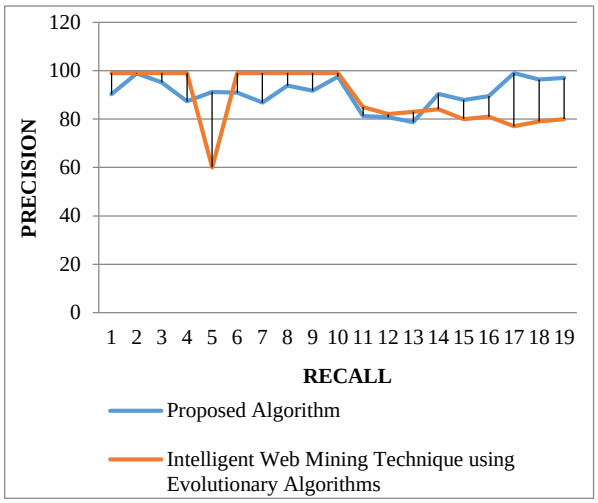

Figure 17. Comparison of F_score values between the proposed algorithm and intelligent web mining technique using evolutionary algorithms [17]

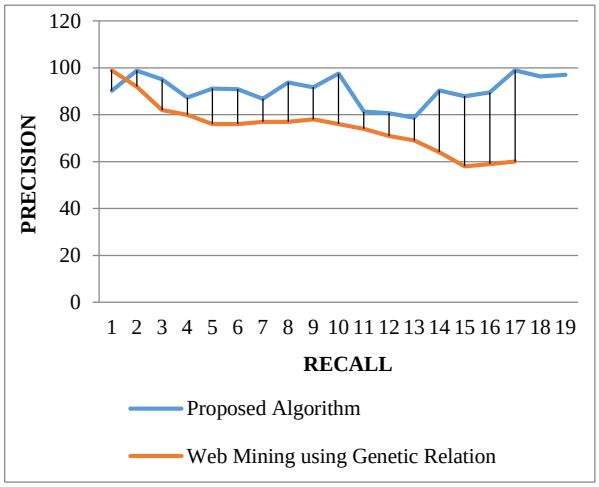

Figure 18. Comparison of F_score values between Proposed Algorithm and Web Mining using Genetic Relation Algorithm [15]

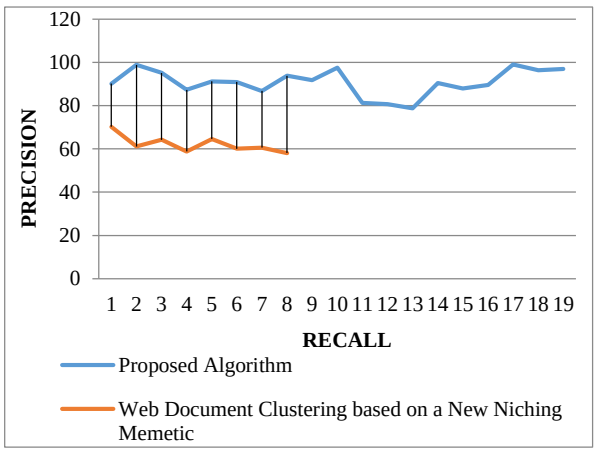

Figure 19. Comparison of F_score value between Proposed Algorithm and Web Document Clustering based on a New Niching Memetic [18]

Table 2. Comparison of average F_scores

Proposed Algorithm	90.7
Intelligent Web Mining Technique using Evolutionary Algorithms [17]	88.5
Web Document Clustering based on a New Niching Memetic [18]	74.6
Web Mining Technique using Genetic Relation Algorithm [15]	62.2

4. Conclusion

The proposed algorithm explores a big search space which was not possible to be searched previously. This is because of new fields which have been used in the proposed algorithm in comparison with related works. In fact, using the basic genetic algorithm and without the proposed categorization, lots of time will be spent on the calculation. Actually, this is the reason for using the method of matrix clustering in the genetic algorithm.

However, it should be noted that there should be appropriate and reasonable link pages between different states of a problem. Finally, genetic algorithm allows to move fast toward the target of the problem, as we are flying towards it. Genetic algorithm are accurate and fast. Genetic algorithm along with the definition of proper and correct functions can provide a suitable and correct answer. Web mining through genetic algorithms is an evolutionary technique for search engines.

In this technique, each generation is improved compared to the previous generation, and the modified population only remains at the end. This remained populations are the proposed pages to the user, which are in the highest grade in terms of quality and similarity. In summary, this method is similar to eugenics, in which a better generation is produced in each breeding until the best is generated, and no best generation is further produced and the qualities of selected pages remain same.

References

[1] Kanya, N. & Geetha, S. (2007). Information Extraction –A Text mining approach ICTES. pp. 1111-1118.

[2] Lee, H. K.,An B. S., Kim, E. J. (2009). ,Adaptive Prefetching Scheme Using Web Log Mining in Cluster-based Web Systems, IEEE International Conference on Web Services(ICWS), IEEE COMPUTER SOCIETY, pp. 903- 910.

[3] Markov, Z. & Larose, D. (2007). Data Mining the Web (Uncoverin Pattern in We Content, Structure, and Usage).

[4] Chen, P., Sun, C., Yang, Y. (2008). Modeling and Analysis the Web Structure Using Stochastic Timed Petri

Nets, journal of software (JSW, ISSN 1796-217X), academy publisher, vol. 3, no. 8, pp.19-26.

[5] Bruno, A. & Pimentel, Renata M.C.R. de Souza, A (2013). Multivariate fuzzy c-means method, Elsevier.com.

[6]Arzanian, B., Moradi, P., Akhlaghian, F., A.(2010) Multi-Agent Based Personalized Meta-Search Engine Using Automatic Fuzzy Concept Networks, Third international conference on knowledge discovery and data mining, pp 208-211.

[7] Yang, S., Chen, P., Sun C., (2007). Using Petri Nets to Model the Web Structure, 4th International Symposium on Applied Computational Intelligence and Informatics (SACI '07), IEEE, pp. 231-235.

[8] Alhalabi, W., Kubat, M., Tapia, M. (2006). Search Engine Personalization Tool Using Linear Vector Algorithm, Proceedings of the 4th Saudi Technical Conference and Exhibition.

[9] Turban, E., Sharda, R., Aronson, J.E., and King, D. (2011) Business Intelligence A managing approach, Second edition, Chapter 5.

[10] Cayzer,S. & Twycross, J. (2003). An Immune-Based Approach to Document Classification, Proceedings of International Conference on Intelligent Information Processing and Web Mining, Zakopane, Poland, pp. 33-48.

[11] Yang, S., Chen, P., Sun C. (2007). Using Petri Nets to Enhance Web Usage Mining, Acta Polytechnica Hungarica, ISSN 1785-8860, vol. 4, no. 3, pp. 113-125.

[12] Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. In SIGKDD Explorations Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2(1), pp 1-15.

[13] Bautista, M. J. M., Amparo, M., Henrik, V., and L. Larsen, (1999). A genetic fuzzy classifier to adaptive user interest profiles with feature selection, In Proc. of the European Society for Fuzzy Logic and Technology (Eusflat-Estylf), Joint Conference, pp. 327-330.

[14] Hossaini, z., Rahmani, A.M., Setayeshi, S., (2008). Web pages Classification and Clustring by means of Genetice Algorithms: A Variable size Page Representing Approach, computational Intelligence for Modeling Control & Automation, 2008 International Conference on Computing & processing (hardware /software) IEEE10.1109/CIMCA. Page(s): 436-440.

[15] Gonzales, E. Mabu, Sh., Taboada, K., and Hirasawa, K. (2010). Web Mining Using Genetic Relation Algorithm. J. of Institute of Electrical Engineers of Japan, Vol. 6, No. 5.

[16] Kumar, R., Tyagy, S., Sharma, M. (2013). Memetic Algorithm: Hybridization of Hill Climbing With Selection Operator, International Journal of Soft Computing and Engineering(IJSCE),2231-2307.

[17] More, S. & Bharambe, U. (2014). Intelligent Web Mining Technique Using Evolutionary Algorithms.978-1-4799-2900-9/14.

[18] Cobos, C., Montealegre, C., Fernanda, M., Mendoza, M., Leon, E., (2010) Web Document Clustering based on a New Niching Memetic Algorithm, Term Document Matrix and Bayesian Information Criterion, 978-1-4244-8126-21101.

Ä f § Z Ë { Â ^ Æ] ® Ì f ¿ ƒ º f Ë • Â ´ · Y • Y Ã { Z ¨ f ‡ Y Z] É Á Z ¯ [Á

[redacted]

ارسال **/**/****؛ پذیرش **/**/****

چکیده

وب جهان گستر با سرعت بسیار زیادی در حال رشد است و اطلاعات بسیار زیادی را در دسترس همگان قرار می دهد.موتورهای جستجو از روش های مرسوم و متداول برای بازیابی اطلاعات در وب استفاده می کنند اما نتایج بدست آمده از این موتورها هنوز قابلیت پالایش را داردو دارای دقت کافی نمی باشد.یکی از روش ها برای وب کاوی الگوریتم های تکاملی می باشد که جستجوها را مطابق با منافع کاربر انجام می دهد.الگوریتم ژنتیک جدید روابط مهم بین پیوندها در صفحات وب را با الگوریتم های تکاملی و خوشه بندی بهینه می کند.روش الگوریتم پیشنهادی مکمل موتورهای جستجو می باشد. در روش پیشنهادی پس از پیاده سازی الگوریتم ژنتیک با استفاده از نرم افزار MATLAB 2013 و نرخ تقابل ۰,۷ و نرخ جهش ۰,۰۵، بهترین و مشابهترین صفحات به کاربر ارائه می شود. در این الگوریتم انتخاب هرکدام از کروموزمها تاثیری در نتیجه ندارد زیرا پس از تولید چندین نسل با نرخ تعیین شده کیفیت کروموزمها یکسان می شود.

کلمات کلیدی الگوریتم ژنتیک، وب کاوی ،محاسبات تکاملی