Web Log Mining Based-on Improved Double-Points Crossover Genetic Algorithm

Jin Xie
College of Computer Science, South-central University for Nationalities, Hubei 430074, China
Email: glady8311@126.com

***Abstract*—Web log files have become important data source for discoveries of user behaviors. Analyzing web log files is one of the significant research fields of web mining. This paper proposes an improved double-points crossover genetic algorithm for mining user access patterns from web log files. Our work contains three different components. First, we design a coding rule according to pre-processed web log data. Second, a fitness function is presented by analyzing user sessions. Finally, a new genetic algorithm based on double-points crossover genetic algorithm is designed. In comparison with simple genetic algorithm, double-points crossover genetic algorithm demonstrates better convergence than the other, and it is more suitable for web log mining. We conducted an experiment to verify the effectiveness of the proposed algorithm. The results show that the proposed algorithm helps the website to easily gain access patterns.**

***Index Terms*—Web Log Mining; Improved Double-point Crossover Genetic Algorithm; Binary Code; User Access Patterns**

I. INTRODUCTION

As the internet is expanding swiftly in the aspects of flowing, dimensions and complexity, WWW has already become a huge knowledge treasury and information ocean in the various walks of society. According to "32 China Internet Development Report" released by CNNIC (the China Internet Network Information Center), by the end of June in 2013, the number of netizens are up to 591 million. Great deals of databases are used widely in various work fields, such as enterprise, bank, government and research institute. In one aspect, technology of network offers technical support for data transmission and remote interaction. In the other aspect, DBMS (Data Base Management System) provides technical assurance and development platform for storage and data management. However, there is still huge gap between data explosion and knowledge deficiency. Meanwhile the explosive growth of information on the web may bring some search problems: (1) general purpose search engines often return too many irrelevant results when users are searching for specific information on a given topic and (2) the number of pages to be indexed by web search systems has been increasing day by day, which makes it difficult to keep both automated and human-maintained indices up to date [1]. Nowadays, internet search engines focus on information retrieval, which can't be used in potential knowledge discovery. There exists a technique which is called Web Mining by connecting KDD (Knowledge Discovery in Database), Data Mining and Web to solve these problems.

Recently, mining of Web information has gained some development and some achievements, especially, in the aspect of Web log mining. V. Sujatha presented the Prediction of User navigation patterns using Clustering and Classification from web log data [2]. Z. Zhao also presented clustering algorithm in data mining based on Web log [3]. Abdelghani Guerbas suggested the usage of a specific density based algorithm for navigational pattern discovery [4]. Yu-Shiang Hung introduced sequential profiles for elder self-care behavior patterns were captured by applying sequence-based representation schemes in association with Markov models and ART2-enhanced K-mean clustering algorithm for sequence behavior mining cluster patterns for the elders [5]. Some studies applied web usage mining directly in cloud computing and website. Such as, Joan M described a distributed algorithm for sequential mining within registry federation in clouding computing [6]. C.J. Carmona presented the methodology used in an e-commerce website of extra virgin olive oil sale called www.OrOliveSur.com [7].

As a result of characters of web, such as diversity of web data, user diversity and network distribution, the technology of Web log mining face some problems. As the following: firstly, it is hard to get the information from mass data. Secondly, it's difficult to mine hidden knowledge from Web data. Thirdly, it's unable to follow the tracks of user interests accurately and lack of informational service personally.

Genetic Algorithm (GA) is a kind of computational model which simulates natural selection from Darwin's theory of evolution and the process of biological evolution from genetic mechanism. This algorithm begins with initial population which probably has solution set. A population consists of certain numbers of individual by encoding which have chromosome character. As the carrier of hereditary materials, chromosome which is group of genes inside decides individual shape and outer expression. Compare to traditional algorithms, it has the following advantages: (1) Special code of parameters as

Project supported by "the Fundamental Research Funds for Central Universities", South-Central University for Nationalities (Grant Number: CZQ13008).

doi:10.4304/jmm.9.6.804-809

operand. (2) GA is used in optimizing among point group. (3) GA utilizes fitness value dispensing with derivative and other auxiliary information. (4) GA is distinguished from other uncertain decision rules by using probability search.

GA as an optimized algorithm has been widely applied to many fields. Such as M. Cheng used genetic algorithm to embed and extract parameters optimization for protect the copyright of digital image [8]. Hiroaki .Nishino proposed immune algorithm to ease the creation of varied 3D models based on genetic algorithm [9]. Bogdan Tomoiaga adopted an original genetic algorithm based on connected graphs to reconfigurate the distribution system [10]. D. Mourtzis designed manufacturing network for mass customization using a genetic algorithm [11] and so on.

GA is also applied to the area of data mining. Emine Tug presented ALMG (Automatic Log Mining via Genetic) to mine web log files via genetic algorithm and to extract information from data which is placed at sever [12]. Babak Sohrabi used genetic algorithm, neural networks, and collaborative filtering to improve e-commerce websites usability [13]. Arben Asllani used genetic algorithm to ensure multiple criteria web-site optimization, and provide dynamic and timely solution [14]. Amelia Zafra introduced a multi- objective grammar based genetic programming algorithm, MOG3P-MI, to solve a Web Mining problem from the perspective of multiple instance learning [15].

As web log mining can discover hidden information from web log file, it is crucial for adjusting the network service and consequently simplifying user operation. Currently some different difficulties have emerged from web log mining techniques, such as being hard to gain the requisite information hardly and being unable to track users' access hotspots accurately, etc. Genetic algorithm is a heuristic and robust search algorithm based on the Darwinian theory of evolution. According to some research, there is similarity between user access behaviors and biological evolution. In GA, evolutionary process is based on encoding which has directly effect on algorithm function and diversity of population. The methods of encoding always have binary string encoding, float number encoding, hybrid encoding. Compare binary string encoding and float number encoding, the former one has more search capability; the latter can keep diversity in mutation. Combine these two encoding method, hybrid encoding is created. In this paper, we use hybrid encoding to design coding rule. In summary, this paper use genetic algorithm to find out frequently-used user pattern through user access behaviors. After the web log file has been pretreated, the standard of binary code is set up with attributions of user access and web pages. Finally, we design suitable genetic operators and fitness function which are used in improved double-points crossover genetic algorithm (IDCGA) for seeking optimum solution. At the end of the work, we report result on experiments to show our algorithm's efficiency.

This paper is organized as follow. Section 2 introduces web mining and genetic algorithm. Pre-processing of web log file and fitness function are also designed in section 2. Double-points crossover genetic algorithm is described in section 3. Meanwhile advantages of this algorithm in this section are summarized by comparing to simple genetic algorithm. Section 4 describes our algorithm in detail. Section 5 presents the experiment results and analysis. In section 6, we conclude our work.

II. Related Work

Web is a virtual society. It is not only about data, information and service, but also about interactions among people, organizations and automatic systems [16]. Data mining by automatic or semiautomatic exploration and analysis on a large amount of data items set in a database can discover potentially significant patterns inherent in the database [17]. Web mining is a process of extracting knowledge and information from web. Generally, Web mining can be divided into three categories: Web content mining, Web structure mining and Web log mining, in which Web log mining means obtaining information from accessing log files, in another word, it mines information from location mode by which user access Web sites [18].

A. Web Log Mining

Web log file include user ID, search time, URL, rank of URL in search result, sequence number and so on. Analyzing these data, user's basic behavior and mutual association will be explored. These provide supports directly for researching user behavior pattern, evaluating performance of website, etc. The results accrued from the mining of web logs can also be used (1) to personalize the presentation of web contents; (2) to improve user navigation; (3) to improve web design or e-commerce sites; and (4) to improve the customers' satisfaction [19].

In general, Web log mining consists of three phases: data pre-processing, mining algorithm and pattern analysis. Fig. 1 shows the basic process of Web log mining.

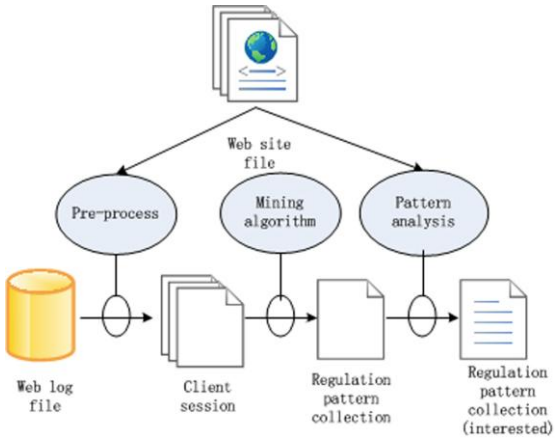

Figure 1. Framework of Web log mining

Recently, Web log pre-processing contains several critical stages: transaction identification, session identification, ID identification, path complement, and data cleaning and so on. Fig. 2 presents a user session and

attributes line from web log file and its descriptions are shown in Table I.

C,"10011",10011 V,1036,1 V,1077,1 V,1003,1 V,1001,1	A,1277,1, "NetShow for PowerPoint", "/stream"

Figure 2. A user session and attribute line

The pattern mining is a kind of technology which can mine pattern and knowledge from pre-processed data. According to requirement analysis, distinct mining technology will be selected. This stage is the core of data mining and also the most difficult point. Generally, pattern mining techniques comprise cluster analysis, classification analysis, association rules, sequential pattern mining and so forth. Pattern analysis evaluates pattern and knowledge which is mined from the second stage through evaluation criterion.

TABLE I. DESCRIPTION OF SESSION AND ATTRIBUTE LINE

User session	Its description
C	Case line mark
10011	User ID
V	Vote line mark
1036,1077,1003,1001	ID of user visited vroots
Attribute line	**Its description**
A	Attribute line mark
1277	Attribute ID number
"NetShow for PowerPoint"	Title of the vroot
"/stream"	The URL relative to root

B. Genetic Algorithm

On the basis of Darwin natural selection theory, GA is a stochastic method by imitating organic evolution. With length of L, n binary strings $b_i(i=1,2,...n)$ form initial solution group. In every binary string, each bit is a gene of individual chromosome. In each generation, new solution population is chose by fitness of individual, by the aid of crossover and mutation of genetic operators.

Fitness function assigns fitness value to each chromosome using genetic structure and relevant information of the chromosome [20].

In this section, our fitness function is described. In the function, session, page weight and length weigh are introduced.

A session is defined that all request of each visitors made to the server in a piece of time [2]. In our dataset, a session concludes case line mark, user ID, vote line mark and ID of user visited vroots. It's easy to distinguish the page number which user visited and their sequence. So in the fitness, means the weight of the visited page. means the length weight of visited sequence.

$$pw_j = \frac{count_j}{COUNT}(j=1,2,...m) \quad (1)$$

$$lw_i = \frac{len_i - \overline{len}}{len_{max} - \overline{len}}(i=1,2,...n) \quad (2)$$

In formula (1) and (2), m is the number of every accessed page, n represents length of accessing in each generation. Then fitness function is designed in this paper as follow:

$$f(i)=\sum_{j=1}^{m} pw_j - lw_i (i=1,2,...n)$$

III. DOUBLE-POINT CROSSOVER GENETIC ALGORITHM

Simple genetic algorithm (SGA) is one point crossover genetic algorithm in which both of parent generations change too much genes. It's easy to destroy excellent individual. However, double-point crossover genetic algorithm is benefit for keeping excellent individual. Fig. 3 describes the flow chart of genetic algorithm.

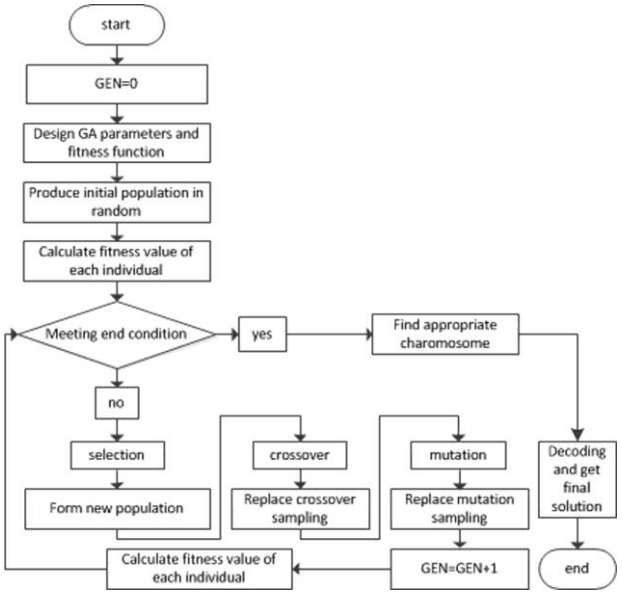

Figure 3. The flow chart of GA

Compare to Fig. 3, double-point crossover genetic algorithm (DCGA) has a little difference in crossover. Crossover is a process of exchange the same bit in different individual for new individual. In DCGA, two crossover points are randomly selected and parent chromosomes are swapped to crossover point (Fig. 4).

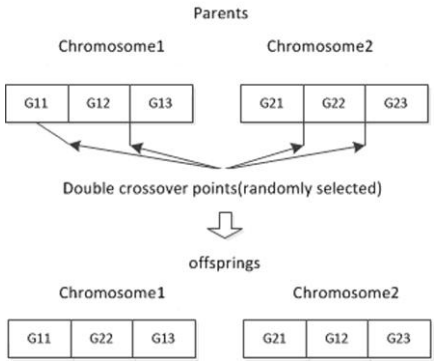

Figure 4. Schematic representation of double crossover point

DCGA demonstrate better convergence than SGA. Fig. 5 shows that the function ($f(x)=(2x+1)e^{-10x^2+2x-1}, 0 \le x \le 1$) was solved by two genetic algorithms.

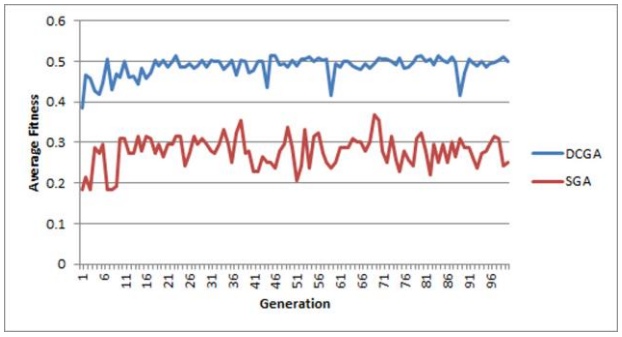

Figure 5. Compare DCGA with SGA

Fig. 6 shows this function value. Comparing Fig. 5 with Fig. 6, the fitness value which was calculated from DCGA was closer to the max value of this function. So DCGA can quickly find the optimal solution.

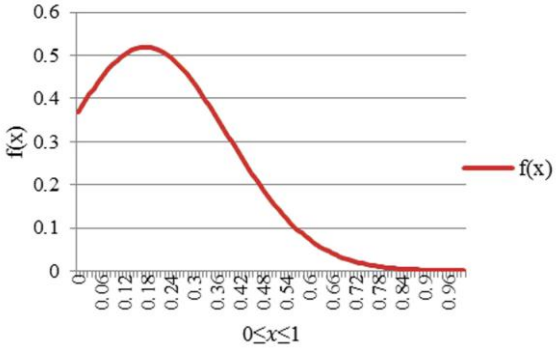

Figure 6. The chart of function

IV. Improved Double-Point Crossover Genetic Algorithm

According to analysis of section 3, this paper selects to ameliorate DCGA to discovery access pattern for keeping efficiency.

Genetic operators include selection operator, crossover operator, and mutation operator.

Selection operator: in accord with sequential number, each individual is given a level i. N is the total number of individuals. So, the probability of selecting the i individual as,

$$P(i)=\frac{2(N-i+1)}{N(N+1)} \tag{4}$$

Crossover and mutation operators: s_1, s_2 as two individuals in crossover operation, s_3 as individual in mutation operation. The maximum value of fitness is

$$f_{max}:f_{max}=\max\{f(s_i),i=1,2,...n\} \tag{5}$$

The average of all fines values is

$$\bar{f}:\bar{f}=\frac{1}{m}\sum_{i=1}^{n}f(s_i) \tag{6}$$

The max of fitness between s_1 and s_2 is

$$f':f'=\max\{f(s_1,s_2)\} \tag{7}$$

P_c is the probability of crossover.

$$P_c=\begin{cases}\dfrac{f_{max}-f'}{f_{max}-\bar{f}}, & f'\geq\bar{f}\\ k_3, & f'<\bar{f}\end{cases} \tag{8}$$

The procedure of program as follow:

Step 1. The initial population is produced randomly. Each individual represents gene code of chromosome. Test whether each of 7 bits is non-repetitive, if not, go to step 1.

Step 2. Every individual fitness value is ascertained by roulette selection and judged by optimization criterion. Success to satisfy these criterions, the best individual and optimum solution are output. If not, turn to step 3.

Step 3. According to fitness value, regeneration of individual is selected. The individual with high fitness value has high opportunity to be selected.

Step 4. In accordance with the probability of crossover and the method of double point crossover, new individual is created. Test whether each of 7 bits is non-repetitive, if not, go to step 4. If step 4 is executed by 10 times, go to step 5.

Step 5. With the probability of mutation and method of mutation, new individual is created. Test whether each of 7 bits is non-repetitive, if not, go to step 5. If step 5 is executed by 10 times, go to step 6.

Step 6. New generational population is created by crossover and mutation. Go to step 2 until the number of generation reach the set number.

V. Experiment and Analysis

Data source come from Microsoft anonymous web data from UCI KDD Archive [21]. The data sources include 5000 anonymous users and 15191 visited records. In this paper, it uses Access 2010 as tool of data pre-processing and Matlab to implement IDCGA.

A. Data Collection and Data Pre-processing

The data was collected by sampling and processing the www.microsoft.com logs. The data records 5000 anonymous, randomly-selected users. Users are identified by a unique sequential number, At first, the extension of data source file is .data. By Access 2010, data source file was transformed to Access database file.

B. Data Analysis

After transforming to database files, it is convenient to analysis. This paper uses VBA in Access to query and summarize the character of data. Fig. 7 shows the number of pages by which every user visited.

In this figure, x coordinate axes stands for users ID, y coordinate axes stands for number of pages. From Fig. 7, it shows that an overwhelming majority of users access below 10 pages. An user who had the most number of accessing visited 28 pages.

Fig. 8 is a distribution curve of each kind number of accessed pages. For example, 318 users access 5 pages.

The ratio is 318/5000(total users number) as equivalent to 0.0636.

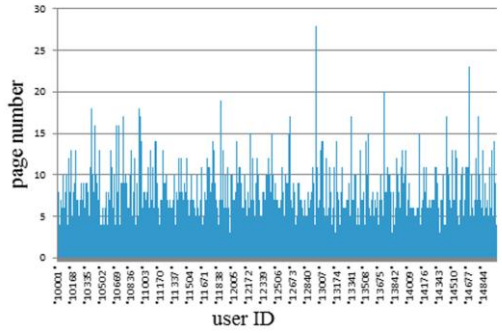

Figure 7. The analysis of user access pages

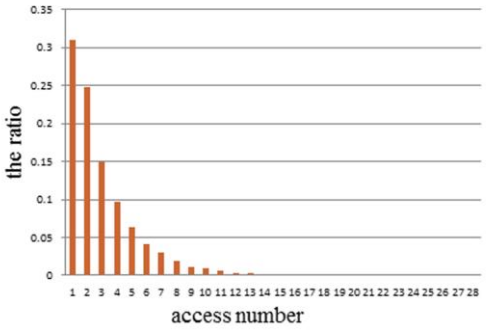

Figure 8. The distribution of access

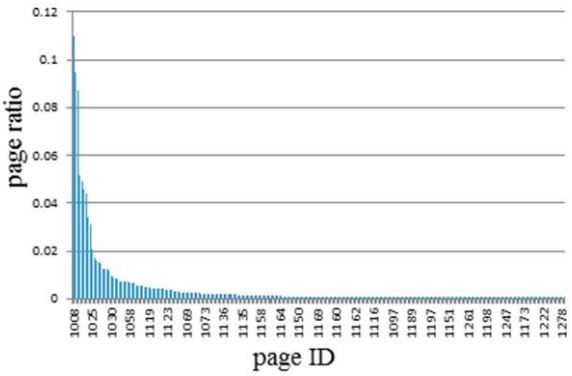

Figure 9. The ratio of each page

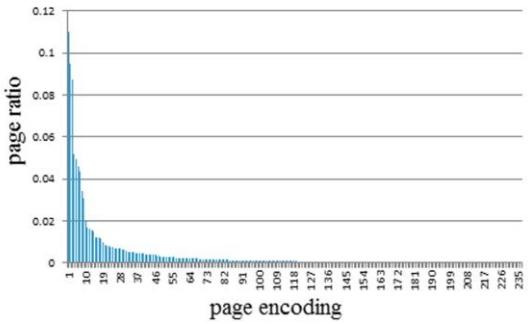

Figure 10. Sequential ID and its ratio

In the data collection, there are 236 pages which were accessed by users. All page access times are calculated and sorted in descending. Fig. 9 shows the ratio of each page access times in total access times. According to this sequence, each page is numbered the unique sequential number, e.g. 1, 2, 3…. Fig. 10 shows sequential number and its ratio.

According to Fig. 8, over 95 percent user's access volumes are less than 7. In accordance with Fig. 10, some pages who's ID after 122 are almost not accessed. So, 122 pages can be encoded by 7 binary bits. If a user visits three pages, the individual has 21 binary bits. If user visits the some page, the binary code is nonzero value. While the number of accessed page is too much, the length of binary code is relative large, it will lower efficiency of system. So that, in this program, the length of binary code is not exceeding 49 binary bits.

C. Discussion

The program is executed in Matlab. And parameters are: the initial number of individual is 30, the number of generation is 20, the probability of crossover is 0.7 and the probability of mutation is 0.03. The length of binary bit is from 7 to 49. Each program execution, the length of binary bit increases 7 bits. E.g.7 bits, 14 bits…49 bits. At result, it can find at 7 patterns from one accessed page to seven visited pages.

After repeated experiments, seven visited patterns are obtained:

http://www.microsoft.com, /msdownload

http://www.microsoft.com, /msdownload, /ie

http://www.microsoft.com, /msdownload, /sitebuilder, /ie

http://www.microsoft.com, /msdownload, /search, /games, /ie

http://www.microsoft.com, /ntworkstation, /ie, /search, /isapi, /regwiz

http://www.microsoft.com, /office, /products, /kids, /msdn, /msdownload, /search

http://www.microsoft.com, /infoserv, /corpinfo, /ie, /msdownload, /powerpoint, /products, /ie_intl

Experimental results are evaluated by three standards [22].

Accuracy= (Total Correct Prediction/Total Predictions)*100%;

Recall= (Total prediction/Total Http Requests)*100%;

AR=Accuracy * Recall.

TABLE II. COMPARE WITH AD

Algorithm	Accuracy	Recall	AR
IDCGA	69.4%	71.3%	49.5%
AD	60%	70%	42%

From Table II, the probability of recall in IDCGA is a little higher than AD and the accuracy is more than AD. So the probability of AR is higher.

VI. CONCLUSIONS

The one of most important research area in Web usage mining is how to mine user behavior patterns in Web log files. In consideration with a large number of data, the paper designs binary code by data analysis. These binary codes are used in improved double-point crossover genetic algorithm. Using IDCGA to find out user

behavior patterns, it is another method for getting correct information in website. At last of this paper, it uses experiment to prove the effectiveness of this method. These results can be used for prediction of user behavior in the website.

REFERENCES

[1] O. SA, A web page classification system based on a genetic algorithm using tagged-terms as features. *Expert Systems with Applications*, Volume 38, Issue 4, April 2011, pp. 3407-3415

[2] V. Sujatha, Punithavalli. Improved user Navigation Pattern Prediction Technique from Web log Data. *Procedia Engineering*, Volume 30, 2012, pp. 92-99

[3] Z. Zhao, Clustering Algorithm in Data Mining Based on Web Log. *Journal of Networks*, Volume 8, 2013, pp. 2392-2399

[4] A. Guerbas, O. Addam, etc. Effective web log mining and online navigational pattern prediction. *Knowledge-Based Systems*, Volume 49, 2013, pp. 50-62

[5] Y. S. Hung, K. L. B. Chen, etc. Web usage mining for analyzing elder self-care behavior patterns. *Expert Systems with Applications*, Volume 40, Issue 2, 1 February 2013, pp. 775-783

[6] Joan M. John, G. Venifa Mini, E. Arun. User Profile Tracking by Web Usage Mining in Cloud Computing. *Procedia Engineering*, Volume 38, 2012, pp. 3270-3277

[7] C. J. Carmona, S. Ramirez-Gallego, F. Torres, etc. Web usage mining to improve the design of an e-commerce website: *OrOliveSur. com. Expert Systems with Application*, Volume 39, Issue 12, 15 September 2012, pp. 11243-11249

[8] M. Cheng, Yan. Li, Y. Zhou, M. Lei, A Combined DWT and DCT Watermarking Scheme Optimized Using Genetic Algorithm. *Journal of Multimedia*, Volume 8, Jun 2013, pp. 299-305

[9] H. Nishino, T. Sueyoshi, T. Kagawa, K. Utsumiya, An Interactive 3D Graphics Modeler Based on Simulated Human Immune System. *Journal of Multimedia*, Volume 3, Jul 2008, pp. 51-60

[10] B. Tomoiaga, M. Chindris, A. Sumper, R. Villafafila-Robles, A. Sudria-Andreu, Distribution system reconfiguration using genetic algorithm based on connected graphs. *Electric Power Systems Research*, Volume 104, November 2013, pp. 216-225

[11] D. Mourtzis, M. Doukas, F. Psarommatis, Manufacturing Network Design for Mass Customization using a Genetic Algorithm and an Intelligent Search Method. *Procedia CIRP*, Volum 7, 2013, pp. 37-42

[12] E. Tug, M. Sakiroglu, A. Arslan, Automatic discovery of the sequential accesses from web log data files via a genetic algorithm. *Knowledge-Based Systems*, Volume 19, Issue 3, July 2006, pp. 180-186

[13] B. Sohrabi, P. Mahmoudian, I. Raeesi, A framework for improving e-commerce websites usability using a hybrid genetic algorithm and neural network system. *Neural Computing and Applications*, Volume 21, Issue 5, July 2012, pp. 1017-1029

[14] A. Asllani, A. Lari. Using genetic algorithm for dynamic and multiple criteria web-site optimizations. *European Journal of Operational Research*, Volume 176, Issue 3, February 2007, pp. 1767-1777

[15] A. Zafra, E. L. Gibaja, S. Ventura. Multiple Instance Learning with Multiple Objective Genetic Programming for Web Mining. *Applied Soft Computing*, Volume 11, Issue 1, January 2011, pp. 93-102

[16] B. Liu, Web Content Mining, the 14th international World Wide Web Conference (WWW -2005)

[17] P. Chou, P. Li, K. Chen, M. Wu. Integrating web mining and neural network for personalized e-commerce automatic service. *Expert Systems with Applications*, Volume 37, Issue 4, April 2010, pp. 2898-2910

[18] A. Budanitsky, G. Hirst, Evaluating Wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, Volume 32, March 2006, pp. 13-47

[19] B. Abedin, B. Sohrabi, Graph theory application and web page ranking for website link structure improvement. *Behaviour & Information Technology*, Volume 28, Issue 1, 2009, pp. 63-72

[20] T. P. Patalia, G. R. Kulkarni, Behavioral analysis of genetic algorithm for function optimization. Computational *Intelligence and computing research (ICCIC), 2010 IEEE International Conference*

[21] http://kdd. ics. uci. edu/

[22] J. Han, H. Zhong, Q. Cai, Prediction for Visiting Path on Web. *Journal of Software*. Vol. 13, No. 6, 2002, pp. 1040-1049