Journal of Theoretical and Applied Information Technology
31st December 2014. Vol.70 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

WEB STRUCTURE MINING FOR USERS BASED ON A HYBRID GA/PSO APPROACH

B. RAJDEEPA[1], DR. P. SUMATHI[2]

[1] Research Scholar, Chikkana Government Arts College, Tirupur & Assistant Professor, Dept. of Computer Science, PSG College of Arts & Science, Coimbatore, India

[2] Assistant Professor, PG & Research Department of Computer Science, Government Arts College, Coimbatore, India

E-Mail: [1] rajdeepab@gmail.com [2] sumathirajes@hotmail.com

ABSTRACT

Web Mining is a demanding task that looks for Web access model, Web structures and the reliability and dynamics of the Web contents. It offers capable Web Personalization, System development, Site alteration, Business Intelligence and Usage Characterization. A latest approach is offered for the estimation of the web site hyperlink structure, where a web user advantage utility function is planned based on known assumptions. The aspire of this work is to propose and get better the Web Structure according to an effectiveness criterion. The frequencies of procedure and time division to click on a link are utilized for constructing the web user utility. Researching probabilistic behaviours have permitted quality development of the hyperlink structure, using a Hybrid GA/PSO. The planned methodology is tested on a real web site and the results are interpreted using the web user benefits by links.

Keywords- *Web mining, web structure mining, web content mining, Hybrid GA/PSO.*

1. INTRODUCTION

Web session clustering is one of the essential Web usage mining techniques which aspire to cluster usage sessions on the basis of a few comparison measures. With the fast growth in Internet technology, the amount of web pages and the number of document content have led to a detonation in the quantity of presented information. While there may be several web pages that are most applicable, popular, or dependable than others, web users look forward to simply, search the most attractive and important website by specifying relevant keywords. It is identified that website appearance is one of the main universal and powerful function on the Internet. As web search engines contain complete range of websites, it helps the search engine to maintain consumers on enormous variety of topics. According to the type of web structural data, web structure mining can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

Web structure mining is also known as "Link Analysis" process. Web mining can be viewed as the extraction of structure from an unlabeled and semi structured data set, as it holds the uniqueness of users detailed documents. It is separated into web content and web usage mining. Web content mining is extraction of the web page content and the environment of transactional database and achieving helpful information from the web text and the explanation of the content information concerning the websites. Web usage mining is completed by mining the suitable website log files and linked data to determine regular browsing patterns based on click stream data analysis. An advanced heuristic intelligent mining can be obtained by using Hybrid GA/PSO for addressing the web mining.

Mahmood and Amjad (2010) believes that the trouble of insertion copies of objects in a distributed web server system will diminish the cost of allocation and read/write requirements,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

when the web servers have restricted storage capacities. They assume the problem as a 0–1 optimization difficulty and present a hybrid particle swarm optimization algorithm to resolve it. The proposed hybrid algorithm makes utilize of the strong global search capacity of Particle Swarm Optimization (PSO) and the strongly built local search ability of tabu search, a search method to achieve high quality solutions. The efficiency of the proposed algorithm is confirmed by comparing it with the Genetic Algorithm (GA), simple PSO, and random assignment algorithm on a mixture of test cases. The simulation results specify that the proposed hybrid approach outperforms the GA and simple PSO.

The remainder of the work is ordered as below. A brief survey is given in Section II. A proposed approach of Hybrid GA/PSO is given in Section III. In Section IV Experimental results for Hybrid GA/PSO are discussed. Conclusions are haggard in Section V.

2. LITERATURE SURVEY

Web Data Mining is an essential area of Data Mining which deals with the extraction of attractive information from the World Wide Web; it is able to categorize into three dissimilar varieties i.e. web content, web structure and web usages mining. The aim of this work is to offer past, present assessment and modernize in each of the three dissimilar variety of web mining are shown above and also outline key future research instructions are done by Singh et al (2010). A novel website structure optimization model is given by Wen-long (2008).

Shutong et al (2009), Site structure optimization method is based on the visitor's browsing performance, optimizing the site arrangement and recording the users visit which helps to predict users visiting site rapidly and correctly in order to diminish the users' access time. Visitors can considerably diminish the "unnecessary" clicks; achieve the target page rapidly, thereby dropping the amount of requirements to Web servers to diminish the burden on the server.

Hussain et al (2010) shows, the web applications are growing at a huge speed and its users are rising at exponential speed. The evolutionary changes in knowledge have made it feasible to capture the users' knowledge and connections with web applications in web server log file. Web log file is saving as(.txt) text file. Due to huge quantity of "irrelevant information" in the web log, the unique log file can not be utilized in the Web Usage Mining (WUM) process. Consequently the preprocessing of web log file becomes very important. Web log preprocessing is initial essential step to get better the excellence and effectiveness of the next steps of WUM.

Particle Swarm Optimization is a population based globalized look for algorithm that mimics the ability (cognitive and social performance) of swarms. PSO manufacture improved results in complex and multi-peak troubles. This effort is useful for the researchers who are effective in the area of PSO and data grouping. PSO variants are also explain in this work are offered by Rana et al (2011).

Yin et al (2013) shows with the fast improvement in World Wide Web (WWW) technology, the amount of webpages and the quantity of information content have been overpowering. It becomes increasingly significant to help users discover appropriate webpage and information more simply and rapidly. This condition causes widespread interest in constructing adaptive websites which mechanically restructure the structure or content by learning from the users' browsing behaviors; as such the usage of the websites is enhanced.

Revelle et al (2010) apply the thought of data fusion to feature location, the procedure of identifying the source code that implements definite functionality in software. A data combination model for characteristic location is offered which defines new feature location methods based on grouping information from textual, dynamic, and web mining analysis applied to software.

Web Usage Mining (WUM) is a kind of Web mining, which utilizes data mining methods to take out precious information from navigation performance of World Wide Web users. The information should be preprocessed to get better the effectiveness and ease of the mining procedure. So it is significant to describe user access patterns from Web log before applying data mining techniques. This work mostly focus on data preprocessing phase of the first phase of Web usage mining with activities like field extraction and data cleaning algorithms given by Aye and Theint (2011).

Akhshabi et al (2014), proposes a PSO algorithm based on Memetic Algorithm (MA) that hybridizes with a restricted look for technique for solving a no-wait flow shop

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

arrangement difficulty. The major purpose is to diminish the whole flow time. In addition, a self-organized chance immigrant structure is expanded into the planned algorithm in order to additional improve its examination capability for original peaks in search space.

In the present study, Particle Swarm Optimization was invoked to meliorate PLS-DA via concurrently selecting the best variable subset as well as the related weights and the best amount of hidden variables in PLS-DA, forming a fresh algorithm named PSO-PLSDA. Marinakis et al (2013), this work established a fresh algorithmic environment stimulated advance that Utilizes a fast Particle Swarm Optimization algorithm with a fresh neighborhood topology for productively solving the Feature Selection Problem (FSP). The planned algorithm for the explanation of the FSP, the Particle Swarm Optimization with Expanding Neighborhood Topology (PSOENT), joins a Particle Swarm Optimization algorithm and the Variable Neighborhood Search (VNS) approach.

Eberhart et al (2001) they focus on the engineering and computer science feature of developments, applications, and resources associated to particle swarm optimization. Lastly, resources associated to particle swarm optimization are scheduled, containing books, Web sites, and software. Alam et al (2008) in this work they explain a fresh web session clustering algorithm that uses particle swarm optimization. They evaluation the existing web usage clustering techniques and suggest swarm intelligence based PSO-clustering algorithm for the grouping of web user sessions.

3. PROPOSED HYBRID GA/PSO METHOD FOR WEB STRUCTURE

The proposed method is more successful in research. The approaches are discussed in this research learning's are

3.1 Genetic Algorithm

The unbelievable enlarge in the amount of information on the World Wide Web has gave rise to subject specific crawling of the Web. During a focused crawling procedure, a mechanical Web page categorization mechanism is needed to decide whether the page being considered is on the subject or not. In this learning, a Genetic Algorithm (GA) based mechanical Web page classification system uses both HTML tags and conditions belong to every tag as classification features and studies optimal classifier from the positive and negative Web pages in the training dataset is residential. Our system classifies Web pages by basically computing relationship among the learned classifier and the fresh Web pages. It has several problems to overcome from that difficulty this work use Hybrid GA/PSO.

3.2 Particle Swarm Optimization

PSO is a calculation method that has been modeled on the biological performance of swarms such as bird flock and fish education. A swarm refers to a gathering of an amount of potential solutions where every possible answer is known as a "particle". In normal PSO technique, every particle is initialized with casual positions Xi and velocities Vi, and a function, f (fitness function) is calculated. The plan of PSO is to discover the particle's position that offers the best evaluation of a known fitness functions using the particles positional categorize as its input values. In a k-dimensional look for space, Xi = (xi1, xi2, xi3,...,xik) and Vi= (vi1, vi2, vi3,...,vik). Position and velocities are familiar, and the function is estimated with the fresh coordinates at each step. In each creation, each particle updates itself continuously by following two great values: the greatest place of the element in its neigh-bourhood (known as local best or individual best position) and the greatest location in the swarm at that instance (known as worldwide best position). After judgment of the beyond values, each particle update its location and velocity as follows:

$$V_{i.k}(t+1) = WV_{i.k}(t) + C_1 r_{1.k}(t)(y_{i.k}(t) - x_{i.k}(t)) + C_2 r_2 K(t)(y'k(t)xi.k(t)) \quad (1)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (2)$$

Where:

vi.k is the speed of the i-th subdivision in the t-th iteration of the k-th measurement; xi.k is the location of the i-th subdivision in the t-th iteration of the k-th measurement; rl and r2 are casual numbers in the gap [0, 1]; cl and c2 are knowledge feature, in common, c1=c2=2; w is the inertia weight factor chosen among (0.1, 0. 9). Equation (1) is utilized to compute the particle's fresh speed according to its previous velocity and the distances of its modern position

Journal of Theoretical and Applied Information Technology
31st December 2014. Vol.70 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

from its individual greatest experience and the group's best understanding. The speed is thus designed based on three contributions:

a) A portion of the previous velocity.
b) The cognitive module which is a purpose of the space of the particle from its individual best location.
c) The social factor which is a purpose of the space of the particle from the most excellent particle establishes thus far (i.e. the top of the personal bests). The particle flies towards a new location according to equation (2). The PSO is typically perform with frequent application of equations (1) and (2) until a particular quantity of iterations have been go beyond or when the speed updates are close to zero over an amount of iterations.

This learn suggest an evolutionary-based grouping algorithm based on a hybrid of Genetic Algorithm (GA) and Particle Swarm Optimization Algorithm (PSOA) for arrange grouping in order to diminish Surface Mount Technology (SMT) setup time. In addition, the model valuation results which use order information offered by a worldwide industrial Personal Computer (PC) producer demonstrate that the proposed algorithm is also better to GA-based and PSOA-based clustering algorithms. Through arrange clustering, training order that belong to the same group together can decrease construction time as well as machine inactive time.

3.3 Hybrid GA/PSO

The main idea of hybrid GA/PSO algorithm is to combine the GA operator into the PSO algorithm.

Step 1: Create early population. Randomly produce M□N first population with binary system. M is the amount of particles in a swarm, and N denotes the length of an individual (particle).

Step 2: Calculate fitness. All the individuals are performed by a fitness function.

Step 3: Execute PSO operators. Every individual modernize its location and speed.

Step 4: Judge Termination. If a modernized individual with fresh fitness cannot assure termination condition,

go to step 5, or else the procedure output the last solution.

Step 5: Carry out GA process.

Step 6: Calculate fitness. This step is the same as step 2.

Step 7: Judge Termination.

Once the extinction condition is met, output the last solution, or else go to step 3. The maximum amount of iterations is measured as the termination criterion.

4. EXPERIMENTAL RESULTS

In this effort implemented the mining technique of Web navigational model for the websites. Incremental database are used here to perform the experimental study for websites. The algorithms developed in this work were implemented in MATLAB.

Table 1:
Accuracy and Execution Time for Proposed Method

Methods	Accuracy (%)	Execution Time (Seconds)
GA	76.4	45
PSO	83.9	31
Hybrid GA/PSO	92.8	19

The table 1 shows the accuracy and execution time for GA, PSO and hybrid GA-PSO. It is clear from the above table the proposed approach hybrid GA-PSO gives improved accuracy than other two approaches.

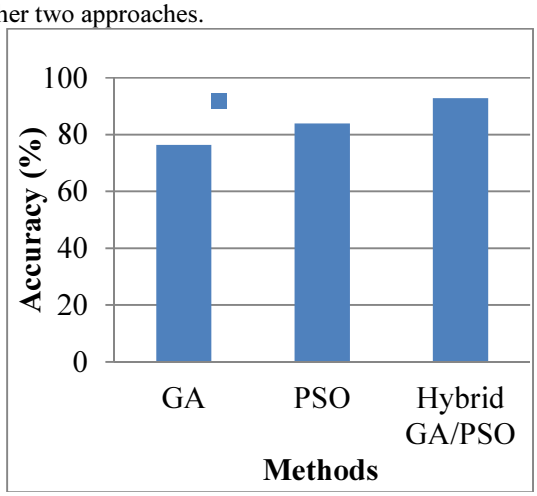

Figure 1: Accuracy For Proposed Hybrid GA/PSO

Figure 1 shows the accuracy for GA, PSO and Hybrid GA/PSO. The proposed method of Hybrid GA/PSO have high accuracy than other two methods.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

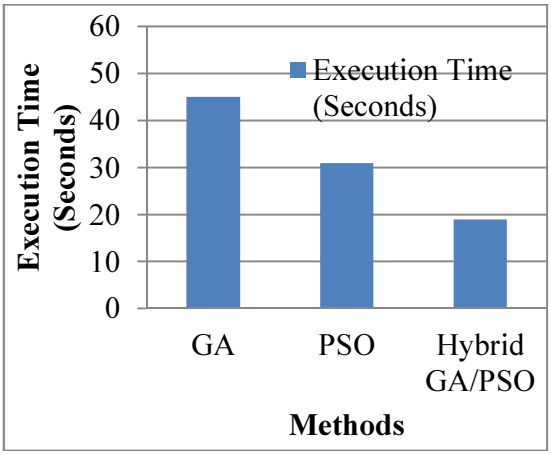

Figure 2: Execution Time for proposed Hybrid GA/PSO

Figure 2 shows the execution time for proposed method of Hybrid GA/PSO. Proposed method has less execution time i.e.19 (sec) when compare with GA and PSO.

5. CONCLUSION

Structure mining uses diminish two main troubles of the World Wide Web due to its huge amount of information. The initial of this difficulty is unrelated search results. Relevance of look for information become misunderstand due to the problem that search engines often only permit for low accuracy criteria. The next of these troubles is the inability to index the huge amount if information offers on the Web. This minimization comes in fraction with the function of discovering the model which is essential for Web hyperlink structure offered by Web structure mining. The simplicity and effectiveness of Particle Swarm Optimization technique based on the idea of Swarm cleverness is being executed in high-dimensional order clustering investigation for web usage mining. As the Hybrid GA/PSO algorithm has fast convergence and simple achievement, numerous enhanced versions of Hybrid GA/PSO algorithm have been developed to resolve that difficulty of GA and PSO.

REFERENCES:

[1] Singh, Brijendra, and Hemant Kumar Singh. "Web data mining research: a survey." In Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on, pp. 1-10. IEEE, 2010.

[2] Hussain, Tasawar, Sohail Asghar, and Nayyer Masood. "Web usage mining: A survey on preprocessing of web log file." In Information and Emerging Technologies (ICIET), 2010 International Conference on, pp. 1-6. IEEE, 2010.

[3] Rana, Sandeep, Sanjay Jasola, and Rajesh Kumar. "A review on particle swarm optimization algorithms and their applications to data clustering." Artificial Intelligence Review 35, no. 3 (2011): 211-222.

[4] Mahmood, Amjad. "Replicating web contents using a hybrid particle swarm optimization." Information processing & management 46, no. 2 (2010): 170-179.

[5] Revelle, Meghan, Bogdan Dit, and Denys Poshyvanyk. "Using data fusion and web mining to support feature location in software." In Program Comprehension (ICPC), 2010 IEEE 18th International Conference on, pp. 14-23. IEEE, 2010.

[6] Yin, Peng-Yeng, and Yi-Ming Guo. "Optimization of multi-criteria website structure based on enhanced tabu search and web usage mining." Applied Mathematics and Computation 219, no. 24 (2013): 11082-11095.

[7] Aye, Theint Theint. "Web log cleaning for mining of web usage patterns." InComputer Research and Development (ICCRD), 2011 3rd International Conference on, vol. 2, pp. 490-494. IEEE, 2011.

[8] Akhshabi, M., Tavakkoli-Moghaddam, R., & Rahnamay-Roodposhti, F. (2014). A hybrid particle swarm optimization algorithm for a no-wait flow shop scheduling problem with the total flow time. The International Journal of Advanced Manufacturing Technology, 70(5-8), 1181-1188.

[9] Li, Y. Q., Liu, Y. F., Song, D. D., Zhou, Y. P., Wang, L., Xu, S., & Cui, Y. F. (2014). Particle swarm optimization-based protocol for partial least-squares discriminant analysis: Application to< sup> 1</sup> H nuclear magnetic resonance analysis of lung cancer metabonomics. Chemometrics and Intelligent Laboratory Systems, 135, 192-200.

[10] Marinakis, Y., & Marinaki, M. (2013). A hybridized particle swarm optimization with expanding neighborhood topology for the feature selection problem. InHybrid Metaheuristics (pp. 37-51). Springer Berlin Heidelberg.

[11] Eberhart, Russell C., and Yuhui Shi. "Particle swarm optimization: developments,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

applications and resources." Evolutionary Computation, 2001. Proceedings of the 2001 Congress on. Vol. 1. IEEE, 2001.

[12] Alam, Shafiq, Gillian Dobbie, and Patricia Riddle. "Particle swarm optimization based clustering of web usage data." Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 03. IEEE Computer Society, 2008.

[13] Shutong Cheng, Congfu Xu.: The Progress of Website Structure Optimization Techniques.J.Application Research of Computers.26 (6), 2013-2015(2009)

[14] LIN Wen-long，LIU Ye—zheng.: A novel website structure optimization model for more effective Web navigation. In: Proceeding of the 1st International Workshop on Knowledge Discovery and Data Mining, pp. 36—41, Adelaide (2008).

[15] Mahmood, Amjad. "Replicating web contents using a hybrid particle swarm optimization." Information processing & management 46, no. 2 (2010): 170-179.