Journal of Information and Operations Management
ISSN: 0976–7754 & E-ISSN: 0976–7762 , Volume 3, Issue 1, 2012, pp-273-276
Available online at http://www.bioinfo.in/contents.php?id=55

SURVEY ON PARTICLE SWARM OPTIMIZATION BASED WEB MINING

STUTI KAROL* AND VEENU MANGAT

Department of Information Technology, University Institute of Engineering and Technology, Panjab University, Chandigarh, India
*Corresponding Author: Email- stuti8karol.oct8@gmail.com

Received: December 12, 2011; Accepted: January 15, 2012

Abstract- Web Mining is a challenging task that searches for Web access patterns, Web structures and the regularity and dynamics of the Web contents. It provides efficient Web Personalization, System Improvement, Site Modification, Business Intelligence and Usage Characterization. High-dimensional Web Log File clustering is a challenging task and requires an efficient clustering technique. The efficiency and simplicity of Particle Swarm Optimization has been exploited for this challenging task and has proved to be a better choice for web session clustering, user profile clustering, page clustering and for many other applications of Web Mining as compared to the traditional K-means clustering method. This paper provides an extensive survey of the application of PSO technique and its variants to Web Usage Mining. Section I of this paper gives a basic introduction to Web Mining, Web Usage Mining and PSO. Section II explains in brief the PSO Clustering technique. Section III discusses in detail the various PSO based techniques used for Web Usage Mining and Section IV concludes the significance of PSO in Web Usage Mining.
Keywords- Data Mining, Clustering Analysis, Web Mining, Web Usage Mining, Web Content Mining, Web Structure Mining, Swarm Intelligence, Particle Swarm Optimization.

Citation: Stuti Karol and Veenu Mangat (2012) Survey On Particle Swarm Optimization Based Web Mining. Journal of Information and Operations Management ISSN: 0976–7754 & E-ISSN: 0976–7762, Volume 3, Issue 1, pp-273-276.

Introduction

Web Mining

Web Mining is a very important discipline of data mining and is drawing huge interest from academia and software industry. The World Wide Web serves as huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services [1][2]. The WWW is expanding tremendously in the number of websites and also in the population of users. This dynamic collection of information provides rich sources for data mining. Web Mining is a challenging task that searches for Web access patterns, Web structures and the regularity and dynamics of Web contents. In Web Mining data can be collected at the server- side, client-side, and proxy servers or obtained from an organization's database [2]. Web Mining can be broadly classified into three classes, i.e. Web Content Mining, Web Usage Mining and Web Structure Mining. Web Usage Mining (WUM) is gaining a lot of interest among researchers [3] as it mines the weblog records to discover user access patterns of Web pages. The activities of large number of internet users generate massive data and provide challenges for the automated discovery of interesting patterns among their usage behaviour. The web users follow some particular sequence while moving from one page or topic to another and discovering such invisible patterns is the ultimate goal of WUM. The analysis of browsing behaviour of web users provides useful guidelines for the improvement of contents, structure, and personalization of web sites.

Web Usage Mining

WUM is basically defined as [4] applying data mining techniques to discover interactions between users and website from web logs. It is well known that users' online interactions with the website are recorded in server web log files that serve as a valuable pool of information. By applying the data mining techniques on web log file, we obtain good insight about the users' behaviour; thereby we can customize the contents and services on the website to better suit the users and we are able to analyse the web logs for various enhancements of website [5]. Fig. (1) [2] shows a Web Log File and Fig. (2) shows the various application areas of WUM. The main three steps in WUM are:

a. *Pre-processing*: it involves converting the usage, content and structure information contained in various available data sources into data abstractions necessary for the purpose of pattern discovery.
b. *Pattern Discovery*: it makes use of methods and algorithms from various fields such as statistics, data mining, pattern recognition and machine learning. There are various mining activities that can be applied to the Web Domain. These activities include Statistical Analysis (used as a tool in Web Traffic Analysis), Association Rules (e.g. association rules refer to the set of pages accessed together with a support value exceeding some specified threshold), Clustering (for developing Usage clusters and Page clusters), Classification (e.g. developing a profile of users belonging to a particular class), Sequential Patterns and Dependency Modelling.
c. *Pattern Analysis*: this is the last step which is used to filter out the uninteresting rules or patterns from the set found in pattern discovery phase.

#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referrer	Agent
1	123.456.78.9	-	[25/Aprl1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Aprl1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Aprl1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Aprl1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Aprl1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
6	123.456.78.9	-	[25/Aprl1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
7	123.456.78.9	-	[25/Aprl1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Aprl1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)

Fig. 1- An example of a Web Server Log

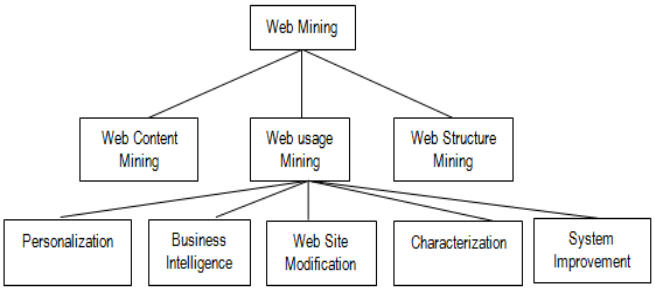

Fig. 2- Application Areas of WUM

Fig. 2- Application Areas of WUM

Particle Swarm Optimization

PSO is a stochastic evolutionary computation technique [6] that has been modelled on the biological behaviour of *swarms* such as bird flocking and fish schooling. A swarm refers to a collection of a number of potential solutions where each potential solution is known as a "particle". In standard PSO method, each particle is initialized with random positions X*i* and velocities V*i*, and a function, *f* (*fitness function*) is evaluated. The aim of PSO is to find the particle's position that gives the best evaluation of a given fitness function using the particle's positional coordinates as input values. In a k-dimensional search space, X*i* = (x_i1, x_i2, x_i3,...,x_ik) and V*i*= (v_i1, v_i2, v_i3,...,v_ik). Positions and velocities are adjusted, and the function is evaluated with the new coordinates at each step. In each generation, each particle updates itself continuously by following two extreme values: the best position of the particle in its neighbourhood (known as *local best* or *personal best* position) and the best position in the *swarm* at that time (known as *global best* position). After finding the above values, each particle updates its position and velocity as follows:

$$v_{i.k}(t+1) = wv_{i.k}(t) + c_1r_{1.k}(t)(y_{i.k}(t) - x_{i.k}(t)) + c_2r_2.k(t)\,(y`k(t)\,xi.k(t)) \quad (1)$$

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

Where:

$v_{i.k}$ is the velocity of the *i-th* particle in the *t-th* iteration of the *k-th* dimension; $x_{i.k}$ is the position of the *-ith* particle in the *t-th* iteration of the *k-th* dimension; r_1 and r_2 are random numbers in the interval [0, 1]; c_1 and c_2 are learning factors, in general, $c_1=c_2=2$; *w* is the *inertia weight* factor selected between (0.1, 0. 9). This parameter was introduced in [7] which illustrated its significance in the Particle Swarm Optimizer. Equation (1) is used to calculate the particle's new velocity according to its previous velocity and the distances of its current position from its own best experience and the group's best experience. The velocity is thus calculated based on three contributions:

a. A fraction of the previous velocity.
b. The cognitive component which is a function of the distance of the particle from its personal best position.
c. The social component which is a function of the distance of the particle from the best particle found thus far (i.e. the best of the personal bests). The particle flies towards a new position according to equation (2). The PSO is usually executed with repeated application of equations (1) and (2) until a specified number of iterations have been exceeded or when the velocity updates are close to zero over a number of iterations.

PSO Clustering

Clustering is a phenomenon to group similar objects based on some common properties, which is called the similarity measure. The elements within a cluster are relatively more similar to each because they have similar properties or attributes. The main objective of clustering is to minimize inter-cluster similarity and maximize intra-cluster similarity. Clustering is one of the most prominent data mining techniques that have been used for various applications such as pattern discovery, data analysis, prediction, visualization, and personalization. Clustering is a common approach to web usage mining for identifying and extracting significant common inter-

ests and similar navigation patterns of web users. Clustering methods can be roughly divided into three categories [8]:

a. *Partitioning methods*: partition data set into *k* clusters. The *k*-means is a typical partitioning based clustering algorithm. It is robust and most popular to cluster the real value data elements in a dataset into *k* groups. Clustering can be done by minimizing the mean square error (MSE) measure referencing to the cluster centroids.
b. *Hierarchical methods*: it does not require the number of clusters to be specified in advance unlike partitioning methods. These methods are deterministic and have lower execution time efficiency than partition based clustering.
c. *Model-based methods*: discover the best fit between data points given a probability distribution.

In the context of clustering [9] [10], a single particle in PSO represents the N_c cluster centroid vectors. That is, each particle *x*, is constructed as follows:

$$x_i = (m_{i1}...m_{ij}..m_{iNc}) \quad (3)$$

Where*:*
m_{ij} refers to the *j-th* cluster centroid vector of the *i-th* particle in cluster C_{ij}. Therefore, a swarm represents a number of candidate clustering for the current data vectors. The fitness of particles is easily measured as the quantization error,

$$J_e = \frac{\sum_{j=1}^{Nc} [\sum_{\forall z_p \in C_{ij}} d(z_p, m_j)/|C_{ij}|]}{N_c} \quad (4)$$

Where *d* is the distance to the centroid given by equation:

$$d(z_p, m_j) = \sqrt{\sum_{k=1}^{Nd} (z_{pk} - m_{jk})^2} \quad (5)$$

and $|C_{ij}|$ is the number of data vectors belonging to cluster C_{ij} i.e. the frequency of that cluster.

The PSO clustering technique is being widely used for the purpose of WUM in terms of Web Session Clustering, Web Usage Data Clustering, creating clusters of user profiles etc. The standard PSO clustering algorithm can stagnate at local optima; hence it will not suit to cluster complex datasets. We need an efficient and flexible algorithm for the Web log data which is high- dimensional, sparse and excessive. The following section discusses variants of PSO technique for Web Usage Mining.

PSO Based Techniques for Web Usage Mining

A few basic PSO techniques that have been implemented over the past few years are discussed below:

A. RVPSO-K Algorithm for clustering Web Usage Pattern

This is K-means Cluster Algorithm based on improved velocity of PSO Clustering Algorithm [11]. The searching ability of the individual particles is enhanced by changing the flying trajectory of the particle, hence the global optimum would be found in a relatively short amount of time. This technique is concluded to be more efficient than the K-means based on PSO clustering algorithm with respect to stability, precision and the convergence speed. Therefore it is quite feasible and effective for clustering web usage pattern.

B. PSO based algorithms for Web Feature Extraction

In [12] a novel approach has been presented that extracts real content from news Web pages in an unsupervised fashion. This method is based on distilling linguistic and structural features from text blocks in HTML pages, having a Particle Swarm Optimizer learn feature thresholds for optimal classification performance. Empirical evaluations and benchmarks show that this approach works very well when applied to several hundreds of news pages from popular media in 5 languages. The novel paradigm presented works in a fully-automated fashion, classifying text blocks within HTML pages as signal or noise by means of linguistic and structural features. Performance evaluations have shown that this approach exhibits an accuracy that comes close to human judgement. A novel Web Feature Extraction Algorithm which is based on the improved particle swarm optimization with *reverse thinking particles* (PSORTP) [13] [14] greatly improves the efficiency of web texts processing. The description of web text has been done through the use of Vector Space Model (VSM) [14]. The trait of the reverse thinking particles is that they do not update their positions, they move to an opposite direction. They are selected randomly in the particle swarm and their new formula which will update their position is:

$$X`_j(t+1) = X_j(t) - V_j(t) \quad (6)$$

When the algorithm starts, normal particles run according to the formula (1) and (2) and the reverse thinking particles run according to the formula (2) and (6). If one of the reverse thinking particles finds a better solution than 'g' (the best position of all particles), the reverse particles will change to a normal particle and one particle will be selected randomly from the normal particles to become a reverse thinking particle. If the 'g' is acquired by the normal particles, the reverse thinking particle will still be opposite. When the reverse thinking particles' position reaches the border of the problem space, they will be set to normal particles which have the best position 'g'. This algorithm can search the multidimensional complex space efficiently. It presents a good idea in future dimension reduction and improving the efficiency for Web Document Processing.

C. Merger of PSO and Agglomerative Algorithm

In Web Usage Mining (WUM), *Web Session Clustering* [2] plays a key role to classify web visitors based on the user click history and similarity measure. Session is defined as a set of user visits to a website in a particular visit and session identification from web log is indeed a complex job. Swarm based web session clustering helps in many ways to manage the web resources effectively such as web personalization, schema modification and website modification. The simplicity and efficiency of PSO technique is of great advantage for the purpose of clustering web usage data. A framework for web session clustering using PSO at the pre-processing level of WUM has been proposed in [4]. In this paper the authors have presented a complete pre-processing methodology of WUM process by applying the data mining techniques. The framework covers the data pre-processing steps to prepare the web log data and convert the categorical web log data into numerical data. A session vector is obtained, so that appropriate similarity and swarm optimization could be applied to cluster the web log data. The authors have implemented a merger of PSO and agglomerative algorithms to obtain hierarchical sessions. The hierarchical cluster based approach enhanced the existing web session techniques for

more structured information about the user sessions.

D. PSO based Sequence Clustering algorithms

Web user session clustering is very important in web usage mining for web personalisation. In [15] the author proposed a partitioning clustering algorithm using a similarity function S^3M which measures sequence and set-similarity, this algorithm using S^3M is more suitable for sequence clustering as compared to the K-means algorithm. Also the author used Total Benefit (TB) measure to estimate "the average intra-cluster similarity with respect to the K-centroids". The use of TB provides better performance to a clustering problem. A proposed swarm intelligence based PSO clustering algorithm [3] for the clustering of web user sessions is made to work independently without hybridization with any other clustering algorithm. The experiments have been performed on the NASA log file [3], and it has been observed that in case of intra cluster distance PSO has performed better than k-means. Also it is observed that more the number of sessions in the cluster, lesser the intra cluster distance and vice versa. The relationship between the number of iterations of PSO clustering and execution time was observed as linear. In [16] the author presents a PSO based sequence clustering approach and an experimental investigation of the PSO based sequence clustering methods which use three original PSO variants and their corresponding variants of a hybrid PSO with real value mutation (PSO-RVM). The investigation is conducted on five web user session datasets extracted from a real world web site. On comparing the experimental results of these methods with the results obtained from the traditional K-means clustering method, it is observed that the PSO and PSO-RVM methods perform better than the K-means method. Furthermore, the PSO-RVM methods show better performance than the corresponding PSO methods in the cases in which the similarity measure function is more complex.

E. PSO based Web Service Selection Optimization Method

The service selection problem in Web service composition deals with the user directly using the functional properties and non-functional properties of Web Service Description Model to search and select the Web service to meet the user's request [17]. Under the service-oriented circumstances, each service may be offered by different providers with different non-functional Quality of Service (QoS) attributes. Since many existing services have similar QoS attributes, the efficiency of Web Service Selection inevitably decreases with the increase of available services. Therefore, a selection process is required to identify which constituent services are to be used to construct a composite service that best meets the QoS requirements of its users. The Particle Swarm Optimization (PSO) algorithm can be used to resolve this multi-objective optimization problem (which is an NP- Complete problem) to resolve dynamic web services selection with global QoS constraints, considering various QoS attributes, such as response time, cost, and availability etc. The PSO-based Web Service Selection method (PSOWSS) to resolve dynamic web services selection with QoS global optimal is shown to be quite feasible and efficient in the process of service selection.

Conclusion

Web Usage Mining (WUM) is currently one of the most interesting topics that are gaining attention from the researchers. WUM is a non-trivial process of extracting useful implicit and previously unknown patterns from the usage of the Web. Significant research is invested to discover these useful patterns to increase profitability of e-commerce sites. Its basic advantage is to provide efficient Web Personalization, System Improvement, Site Modification, Business Intelligence and Usage Characterization. The simplicity and efficiency of Particle Swarm Optimization method based upon the concept of Swarm Intelligence is being implemented in high-dimensional sequence clustering analysis for web usage mining. As the PSO algorithm has fast convergence and easy implementation, many improved versions of PSO algorithm have been developed to resolve that problem of clustering high-dimensional web log sessions. PSO has a wider scope in future in mining the Web and a lot of research is still being done in this challenging field.

References

[1] Yanchun Zhang and Guandong Xu (2008) *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*.

[2] Jaideep Srivastava, Robert Clooney, Mukund Deshpande and Pang Ning Tan. *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,* Vol. 1, Issue 2.

[3] Shafiq Alam, Gillian Dobbie and Patricia Riddle (2008) *International Conference on Web Intelligence and Intelligent Agent Technology*.

[4] Tasawar Hussain, Sohail Asghar and Simon Fong. *A Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining.*

[5] Osmar R. Zaïane. *Web Usage Mining for a Better Web-Based Learning Environment*.

[6] James Kennedy and Russel Eberhart (1995) *International Conference on Neural Networks, Perth, Australia*, pp.1942–1948.

[7] Shi Y. and Eberhart R.C. (1998) *International Conference on Evolutionary Computation*, Anchorage, Alaska.

[8] Shafiq Alam, Gillian Dobbie, Patricia Riddle and Asif Naeem M. (2010) *International Conference on Web Intelligence and Intelligent Agent Technology*.

[9] Van Der Merwe D.W. and Engelbrecht A.P (2003) *Congress on Evolutionary Computation*, Canberra, Australia.

[10] Ching-Yi Cheo and Fun Ye (2004) *international Conference on Networking*, Sensing Control, Taiwan.

[11] Junyan chen and Huiying Zhang (2007) *IEEE*.

[12] Cai-Nicolas, Ziegler Michal and Skubacz Siemens AG. Corporate Research & Technologies Otto-Hahn-Ring 6, D-81730 München.

[13] Zhang Xiaoming and Wang Rujing (2006) Computer Science, 33 (10) : 156~159.

[14] Song Liangtutt and Zhang Xiaoming (2007) IJCSNS, V.7 No.6.

[15] Pradeep Kumar, Raju S. Bapi and Radha Krishna P (2007) International Journal of Data Warehousing and Mining, Vol. 3, Issue 1.

[16] Thi Thanh Sang Nguyen and Haiyan Lu (2009) *International Conference of Soft Computing and Pattern Recognition*.

[17] Yaning Huang and Fei Li (2010) *International Conference On Computer Design And Appliations*.