Research Journal of Applied Sciences, Engineering and Technology 7(7): 1469-1473, 2014
DOI:10.19026/rjaset.7.421
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: September 28, 2013 Accepted: October 14, 2013 Published: February 20, 2014

Research Article

An Heighten PSO-K-harmonic Mean Based Pattern Recognition in User Navigation

R. Gobinath and M. Hemalatha
Department of Computer Science, Karpagam University, Coimbatore, India

Abstract: The website navigation patterns can be searched and analyzed with the introduction of the new methodology. The user navigation path is stored as a sequence of URL categories in web server. The approaches followed are to separate the users and sessions from the web log files and acquiring the necessary patterns for web personalization. The clustering concept is used for grouping the necessary patterns in separate groups. The approaches used for clustering of navigation patterns are done with improvised particle swarm optimization technique which divides users depends on the order in which they request web pages. This approach mines the web log files which are resultant from the web users while interacting with web pages for a particular period of web sessions. The work carried with an optimized method of particle swarm optimization-K-Harmonic means to cluster the similar users based on their navigation pattern. Particle swarm optimization-K-Harmonic method is used to discover or extract user's navigational patterns from web log files.

Keywords: K-harmonic means, particle swarm optimization, web mining

INTRODUCTION

Web mining is the technique of data mining for acquiring knowledge from large databases. The extraction of necessary access logs information for web personalization process undergoes many processes which coincide with the general data mining flow of preprocessing, identifying patterns and analyzing collected patterns. The contribution of many researchers in web mining shows the general three categorization of web mining known as structure mining, content mining and usage mining. The general cleaning of unwanted data from web access log files are completed in the preprocessing stage. The preprocessing stage removes such unnecessary data like multimedia files, duplicate entries, successful status code, etc. from web access log entries and may help in speeding the execution time of the final web personalization process. The web access log files attain a new format after the preprocessing stage which is more suitable for identifying patterns. The identifying patterns from clean web log files follows several methods and algorithm developed from the fields such as data mining, machine learning, pattern recognition and statistics. Analyzing the patterns is the final stage of every web usage mining process and it is possible only after generating the specific rules by applying algorithms in identifying the patterns.

The grouping of necessary patterns should be done, which can be proceeded with the technique known as clustering. The formation of grouping is done by applying clustering algorithms. Researchers have shown their interest in using many kinds of techniques such as k-mean, c-mean and ant based clustering in the pattern discovery process (Nicolas *et al*., 2003). The technique followed for the clustering process in the proposed framework follows some of clustering techniques in common and differ in introducing Particle swarm optimization-K-Harmonic technique for classifying the patterns.

LITERATURE REVIEW

The needed patterns for analyzing the behavior of users can be extracted using web usage mining technique, which is an application of data mining. The research field which deals with the website users's behavior, click stream of the users and session of the user which can be broadly studied in web usage mining (Kay *et al*., 2006). The artificial intelligence came across many new techniques and methods for analyzing the extracted user behavior of the website user's (Cooley *et al*., 1997). The technique of filtering web log files and constructing the sessions from the web log files not only made a change in web usage mining research strategy, it also motivates the researchers to concentrate on understanding the behavior of the website users and collecting the interest of the user from the web servers. The introduction of webMiner system by Cooley *et al*. (1999) helps in extracting the

Corresponding Author: R. Gobinath, Department of Computer Science, Karpagam University, Coimbatore, India

web log files and to build up the web sessions depending upon the transaction patterns in the log files for website expert convenience way of extracting the meaningful transactions from the logs. The webWUM system by Spiliopoulou and Faulstich (1998) which is also very similar to that of the Cooley system webMiner in introducing the SQL-like language for requesting behavior rules. The same system has been developed in 1999 by Spiliopoulou *et al*. (1999) for differencating Profiles from the time of entry and regular visitor. The extraction of ordered patterns and maximizing ISEWUM method can be done by the use of the WebTool proposed by Masseglia *et al*. (1999a, b). The focus of the system is to rearrange the website or web page in a convenient manner for the user. The prediction of web pages by Davison technique (Davison, 2002, 1999) helps to reorganize the website according to the user requirement, Perkowitz and Etzioni (1999) uses a PageGather conceptual clustering algorithm for reorganizing websites by pages index. The clustering methods idea to group the necessary patterns from the web log files and popularly used from early 1996 (Yan *et al*., 1996). The method of clustering is for grouping similar sessions from the log files, which will be very easy method to identify each and every group for further process of identifying the behavior of the user. The access patterns from the log files are identified by the clustering methods for separating the web users (Fu *et al*., 1999). An AntClust system for grouping the web users from the log files are the concept worked by applying the ant clustering algorithm for separating users from the log files (Labroche *et al*., 2003). The Leader Clustering algorithm for grouping the sessions of the log files are used by Yan *et al*. (1996). The method also describes numbers of objects gathered during the active session. AntClust is the clustering algorithm described by Labroche *et al*. (2003), the concept of identifying the data from the data set are very similar in ant colony for identifying the correct path. Labroche (2006) Leader Ant algorithm performs much better than the AntClust algorithm. The Mobasher (2006) approaches for mining user profiles in a web personalization process gives a summary for data mining in the web personalization process.

OVERVIEW OF FRAMEWORK

In our proposed work after collecting the dataset from open repository the data has to be converted from raw format to standard format. The irrelevant and unwanted entries are removed from the dataset in data cleaning phase. The session and user are identified with the help of IP address and time stamp. In the next phase pages are extracted based on their navigation feature. The final phase is discovering similar navigation pattern using our proposed method particle swarm optimization using K harmonic means which improves the traditional particle swarm optimization instead of starting with random cluster centroids here we used K harmonic mean for finding cluster centroid (Fig. 1).

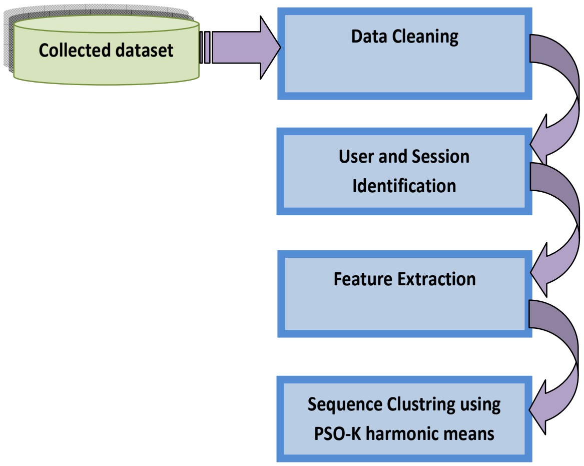
Collected dataset
Data Cleaning
User and Session Identification
Feature Extraction
Sequence Clustring using PSO-K harmonic means

Fig. 1: Structure of the proposed framework

The navigation pattern of a web user often replicates user's psychological model and for this motivation, websites owners and developers pay more attention to the navigation patterns. In sort to study these patterns efficiently, browsing patterns collected from a site is very important. Numerous clarifications have been anticipated and the treatment of clustering and classification is more commonly used on such solutions. This research work is another attempt made to propose a Bio-inspirational based system that uses K harmonic mean based particle swarm optimization clustering method to discover the user's navigation pattern and analyze them from the server's web log file.

METHODOLOGY

Particle swarm optimization: Particle Swarm Optimization (PSO) is a population-based stochastic optimization technique inspired by bird flocking and fish schooling which was originally designed and introduced by Kennedy and Eberhart (1995) and is based on iterations/generations. The algorithmic flowin PSO starts with a population of particles whose positions represent the potential solutions for the studied problem and velocities are randomly initialized in the search space. In each iteration, the search for optimal position is performed by updating the particle velocities and positions. Also in each iteration, the fitness value of each particle's position is determined using a fitness function. The velocity of each particle is updated using two best positions, namely personal best position and global best position. The personal best position, pbest, is the best position the particle has visited and is the best position the swarm has visited since GEST the first time step. A particle's velocity and position are updated as follows:

Table 1: The sample user sequence

User	Sequence								
1	Front page	Front page							
2	News								
3	Tech	News	News	Local	News	News	News	Tech	Tech
4	Msn- News	Business							
5	MSN-sports	Sports	MSN-sports						
6	On-air	Sports							
7	Front page	Summary	On-air	On-air					
8	Front page	Misc	Misc	Front page					
9	News	Travel	Living						
10	News	Weather	Weather	Weather	Weather				

$$V(t+1) = w.v(t) + c_1r_1\big(pbest(t) - X(t)\big) + c_2r_2\big(gbest(t) - X(t)\big)$$

$$k = 2, 3, ..p \qquad (1)$$

$$X(t+1) = X\big(t0 + V(t+1)\big) \qquad (2)$$

where, X and V are position and velocity of particle, respectively. w is inertia weight, c1 and c2 are positive constants, called acceleration coefficients which control the influence of pbest and gbest on the search process, P is the number of particles in the swarm, r1 and r2 are random values in range [0, 1].

K-harmonic means: The *k*-Harmonic Means algorithm (KHM) is a method similar to KM that arises from a different objective function (Bin *et al*., 1999). The KHM objective function uses the harmonic mean of the distance from each data point to all centers:

$$KHM(X, C) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^p}}$$

Here, p is an input parameter and typically $p \geq 2$. The harmonic mean gives a good (low) score for each data point when that data point is close to any one center. This is a property of the harmonic mean; it is similar to the minimum function used by KM, but it is a smooth differentiable function.

Proposed methodology of PSO-K harmonic means:

- In the context of PSO-K harmonic means clustering, before initializing the particles, the data points are randomly assigned to K clusters first.
- Particle fitness is evaluated based on clustering criteria:

$$F(i) = \sum_{i\ =\ 1}^{n} \frac{k}{\sum_{j\ =\ 1}^{k} \frac{1}{\|x_i - c_j\|^p}}$$

where,
p = An input parameter and typically p> = 2

The fitness functions of the particle i, between a data point $X_{i\ =\ 1..n}$ and the cluster center $Cj_{\ =\ 1..k}$, the harmonic mean gives a score for a each data point, when that data point is closed to any one center.

- The velocity and position of the particles are modified using the following equations:

$$V(t+1) = w.v(t) + c_1r_1\big(pbest(t) - X(t)\big) + c_2r_2\big(gbest(t) - X(t)\big)$$
$$X(t+1) = X(t) + V(t+1)$$

where,
V, X = Velocity and positions of particles
w = For inertia weight
c_1, c_2 = Positive constant called acceleration coefficient which controls influence of pbest and gbest on the search process
P = Represent no of particle in the swarm
r_1, r_2 = Random value ranges from [0, 1]

- The new generation is optimized by K harmonic means as given below:
  - The data set is reassigned to clusters according to nearest rule.
  - Cluster centroids, fitness value are recalculated and positions are updated.
- If the position is satisfactory or the maximum iteration is reached, the process is stopped. Otherwise, return to particle fitness calculation stage.

In the sample user navigation sequence we have displayed 10 random user navigation likewise we have used 3000 records for clustering same pattern of user navigation. Initially, k random number of users are consider as initial cluster centers. Then by applying k harmonic mean, the users are assigned to their nearest clusters using the mean distance measure. Next the particle swarm optimization is applied on each cluster and the velocity and position of each particle is updated. Each user object is updated based on the fitness function of particle swarm optimization. The process is continued until the maximum iteration or the user objects are in accurate clusters (Table 1).

EXPERIMENTAL RESULTS

In this study for conducting experiment we used three different data set they are Kdlog (dataset 1), online shopping (dataset 2), msnbc.com (dataset 3) files. The performance study of our proposed method PSO-K-Harmonic mean is compared with two different existing techniques namely Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO). The metrics used for comparison are accuracy, precision, recall, f-measure and time taken.

Table 2 refers to performance of proposed method based on accuracy is shown. The accuracy is calculated using the equation:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

TP = Number of true positives
TN = Number of true negatives
FP = False positives
FN = False negatives

From the Table 3 it is shown that the time taken by the proposed work is very less while comparing the other two existing techniques. The PSO takes highest computation time.

Precision or Confidence (as it is called in Data Mining) denotes the proportion of Predicted Positive cases that are correctly Real Positives. However analogously called True Positive Accuracy (tpa), a measure of accuracy of Predicted Positives.

In contrast with the rate of discovery of Real Positives (taper):

$$precision = \frac{No.of\ Relevant\ pages\ Predicted}{Total\ no.of\ pages\ predicted}$$

From the Table 4 it is shown that the proposed work has highest precision rate 0.87, 0.91 and 0.93 for dataset 1, dataset 2 and dataset 3, respectively.

Recall or Sensitivity (as it is called in Psychology) is the proportion of Real Positive cases that are correctly Predicted Positive (Table 5). This measures the Coverage of the Real Positive cases by the +P (Predicted Positive) rule. Its desirable feature is that it reflects how many of the relevant cases the +P rule picks up:

$$recall = \frac{No.of\ relevant\ pages\ Predicted}{Total\ no.of\ relevant\ pages\ predicted}$$

From the Table 5, the proposed work has highest recall rate 0.78, 0.80 and 0.82 for dataset 1, dataset 2 and dataset 3, respectively.

F-measure is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct results divided by the number of all returned results and r is the number of correct results divided by the number of results that should have been returned. The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0:

Table 2: Performance analysis of proposed work with existing approaches based on accuracy

Techniques	Data set 1	Data set 2	Data set 3
ACO	86.03	89.75	91.26
PSO	90.08	91.14	93.09
PSO-K harmonic means	92.74	94.81	95.96

Table 3: Performance analysis of proposed work with existing approaches based on time taken

Techniques	Data set 1	Data set 2	Data set 3
ACO	20	18	45
PSO	22	20	52
PSO-K harmonic means	19	18	40

Table 4: Performance analysis of proposed work with existing approaches based on precision

Techniques	Data set 1	Data set 2	Data set 3
ACO	0.62	0.71	0.79
PSO	0.83	0.78	0.81
PSO-K harmonic means	0.87	0.91	0.93

Table 5: Performance analysis of proposed work with existing approaches based on recall

Techniques	Data set 1	Data set 2	Data set 3
ACO	0.65	0.68	0.71
PSO	0.69	0.72	0.75
PSO-K harmonic means	0.78	0.80	0.82

Table 6: Performance analysis of proposed work with existing approaches based on F measure

Techniques	Data set 1	Data set 2	Data set 3
ACO	0.63	0.69	0.75
PSO	0.75	0.75	0.78
PSO-K harmonic means	0.82	0.85	0.87

$$F - measure = 2\frac{precision*recall}{Precision+recall}$$

From the Table 6 it is shown that the proposed work has highest F-measure rate 0.82, 0.85 and 0.87 for dataset 1, dataset 2 and dataset 3, respectively.

CONCLUSION

This study uses improvised particle swarm optimization algorithm for clustering. Finally, the prediction based on the clustering result is performed by means of using the K harmonic means which has the better capability of better prediction than other conventional K means technique. The experimental result shows that the proposed technique has better accuracy, Precision, Recall and F-measure of predictive metrics which also shows the clustering performance evaluation.

REFERENCES

Bin, Z., H. Meichun and D. Umeshwar, 1999. K-Harmonic Means: A Data Clustering Algorithm. Hewlett-Packard Research Laboratory, pp: 1-25.

Cooley, R., B. Mobasher and J. Srivastava, 1999. Data preparation for mining world wide web browsing patterns. Knowl. Inform. Syst., 1(1): 5-32.

Cooley, R., J. Srivastava and B. Mobasher, 1997. Web mining: Information and pattern discovery on the world wide web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97).

Davison, B.D., 1999. Adaptive web prefetching. Proceedings of the 2nd Workshop on Adaptive Systems and User Modeling on the WWW. Toronto, pp: 105-106.

Davison, B.D., 2002. Predicting web actions from HTML content. Proceedings of the 13th ACM Conference on Hypertext and Hypermedia (HT'02). College Park, MD, pp: 159-168.

Fu, Y., K. Sandhu and M. Shih, 1999. Clustering of Web users based on access patterns. Proceedings of the 1999 KDD Workshop on Web Mining. San Diego, CA.

Kay, J., N. Maisonneuve, K. Yacef and O. Zaiane, 2006. Mining patterns of events in students' teamwork data. Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006), pp: 45-52.

Kennedy, J. and R.C. Eberhart, 1995. Particle swarm optimization. Proceedings of the IEEE International Conference on Neural Networks. Piscataway, NJ, 4: 1942-1948.

Labroche, N., 2006. Fast ant-inspired clustering algorithm for web usage mining. Proceedings of the Information Processing and Management of Uncertainty Conference. Paris, France, pp: 2668-2675.

Labroche, N., N. Monmarché and G. Venturini, 2003. AntClust: Ant clustering and web usage mining. Proceedings of the Genetic and Evolutionary Computation Conference (Gecco 2003). Chicago, IL.

Masseglia, F., P. Poncelet and M. Teisseire, 1999b. Using data mining techniques on web access logs to dynamically improve hypertext structure. ACM SigWeb. NewsLett., 8(3): 1-19.

Masseglia, F., P. Poncelet and R. Cicchetti, 1999a. Web tool: An integrated framework for data mining. Lect. Notes Comput. Sc., 1677: 892-901.

Mobasher, B., 2006. Data Mining for Personalization. In: Brusilovsky, P., A. Kobsa and W. Nejdl (Eds.), the Adaptive Web: Methods and Strategies of Web Personalization. Springer-Verlag, Berlin, Heidelberg, Vol. 4321.

Nicolas, L., M. Nicolas and V. Gilles, 2003. AntClust: Ant clustering and web usage mining. Proceeding of the Genetic and Evolutionary Computation Conference (GECCO 2003. Springer-Verlag, Berlin, Heidelberg, LNCS 2723, pp: 25-36.

Perkowitz, M. and O. Etzioni, 1999. Adaptive web sites: Conceptual cluster mining. Proceeding of the 16th International Joint Conference on Articial Intelligence. Stockholm, Sweden, pp: 264-269.

Spiliopoulou, M. and L.C. Faulstich, 1998. WUM: A web utilization miner. Proceeding of the Workshop on the Web and Data Bases (WebDB98), pp: 109-115.

Spiliopoulou, M., C. Pohle and L.C. Faulstich, 1999. Improving the effectiveness of a web site with web usage mining. Proceedings of the WebKDD Conference, pp: 142-162.

Yan, T.W., M. Jacobsen, H. Garcia-Molina and U. Dayal, 1996. From user access patterns to dynamic hypertext linking. Proceeding of 5th International World Wide Wibe Conference on Computer Networks and ISDN System. Paris, France, pp: 1007-1014.