International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 04 | Apr -2017 www.irjet.net p-ISSN: 2395-0072

# A Novel Preprocessing Method for Web Usage Mining based on Hierarchical Clustering

**R. Padmapriya[1]**
[1]*Research Scholar, Department of Computer Science, Rathnavel Subramaniam College of Arts & Science, Sulur*
**Dr. D. Maheswari[2]**
[2]*Assistant Professor & HoD- Research School of Computer Studies , Rathnavel Subramaniam College of Arts & Science, Sulur*

---

## Abstract

Web Usage Mining is the method of implementing data mining procedures to extract usage pattern from Web Log files data. There are three phases in Web usage mining - preprocessing, pattern discovery and pattern analysis. There are several preprocessing tasks that must be performed prior to data collected from server log data mining algorithms to apply. This serves to define the value of specific clients, cross marketing strategies across products and the effectiveness of promotional efforts, and so on. Data preprocessing is a data mining technique which involves the transforming of raw data into an understandable format. Data preprocessing is important to insure the ability of web log mining. Result of preprocessing has direct influence on the choosing of mining algorithm. In this research, data preprocessing algorithms are discussed in database-driven applications such as customer relationship management and rule based applications. The preprocessed Web Log File can be suitable for the discovery and analysis of useful information referred to as web mining. Preprocessing may be needed to make data more suitable for data mining. This research summarizes the efficient and complete preprocessing results before actual mining can be performed.

***Keywords:*** Web usage mining, Data preprocessing, Framework, PSO with Hierarchical clustering

## 1. INTRODUCTION

Referable to the increasing popularity of e-commercialism in our daily lives, credit card usages have increased over the years. There are various applications for examining user navigational pattern which uses web usage mining. Web usage mining is technique of web mining. Preprocessing performs a series of processing of web log file covering data cleaning, user identification, session identification, course completion and transaction identification. . This process deals with logging of the data; performing accuracy check; putting the data together from disparate sources; transforming the data into a session file; and finally structuring the data as per the input requirements. Preprocessing phase is a set of interconnected, coherent, and integrated techniques, employed in a succession to create clean and clear results.

Data preprocessing is needed and important phase in web usage mining. The web log file is the data source for data preprocessing method. The aim of data cleanup is to get rid of irrelevant items. The task of User identification is to identify who access the web site and which pages are accessed in the web site. Current research is on data preprocessing methods which are data cleaning and user identification. A different technique is provided for data cleaning, but still there are problems remain in data collection and accuracy metric of user identification. This paper offers a review on algorithm and different techniques utilized in data preprocessing that are applied for web usage mining. Data preprocessing is used for clean the data so that when it provide for the pattern discovery it will distinguish the technique which will used to discover the users' navigational pattern and after treating it will communicate that to pattern analysis so that it will contain only relevant pattern and removes irrelevant pattern.

Web usage mining refers to the automatic discovery and analysis of patterns in click stream and related data collected or generated as an outcome of user interactions with Web resources on single or more Web sites. The main aim of this is to confine, model and examine the behavioral pattern and profiles of user interacting with Web site. The observed patterns are normally interpreted as collections of pages, objects, or re-sources that are often accessed by groups of users with common needs or interests. Sticking with the stock data mining process the overall Web usage mining process can be split into three interdependent stages: data collection and pre-processing, pattern discovery, and pattern analysis.

This remaining paper describes Literature Survey in Section II, Preprocessing methodology is discussed in Section III, Experiments and achieved results in Section IV. Finally, Conclusion of this work is given in Section V.

## 2. LITERATURE SURVEY

Different data mining techniques can be used on web usage data to extract user access patterns and this knowledge can be applied in a diversity of applications such

as system improvement, web site modification, business news etc. For discovering patterns, data abstraction is required by web usage mining. This data abstraction is achieved by Tyagi et al (2010) through data preprocessing. Thorleuchter, Dirk, and Dirk Van den Poel (2012) evaluated concerning companies' success in the business-to- consumer (B2C) environment where consumers choose their preferred e-commerce websites based on these success factors e.g. website content quality, website interaction, and website customization.

A brief overview of various data mining techniques for discovering patterns, and pattern analysis are discussed by Chitraa *et al* (2010). Finally a glimpse of various applications of web usage mining is also offered. A novel web usage mining approach, established along the sequence mining technique applied by Chou et al (2010) to user's navigation behavior, to see patterns in the navigation of internet sites. Mei *et al* (2010) introduced the concept of Web mining, and describing the process of Web data mining in detail: source data collection, data pre-processing, pattern discovery and pattern analysis, using a detailed case of Web mining application in e-commerce. The method proposed by Dimopoulos *et al* (2010) has the advantage that it demands a constant amount of computational effort per one user's action and consumes a relatively small amount of extra memory space.

JADE is quite easy to learn and use. Moreover, it supports many agent approaches such as agent communication, protocol, behavior and ontology. This framework has been experimented by Kularbphettong *et al* (2010) and evaluated in the realization of a simple, but realistic. Pamnani *et al* (2010) discusses an application of WUM, an online Recommender System that dynamically generates links to pages that have not yet been visited by a user and might be of his potential interest. Singh Brijendra and Hemant Kumar Singh (2010) provided past, current evaluation and update in each of the three different types of web mining i.e. web content mining, web structure mining and web usages mining and also outlines key future research directions.

In order to create suitable target data, the further essential tasks of pre-processing Data Cleaning, User Identification, Sessionization and Path Completion are designed by Rao *et al* (2010) collectively. The framework reduces the error rate and improves significant learning performance of the algorithm. Agent-based modeling is used to simulate e-commerce transaction networks. For real-world analysis, Piao *et al* (2010) studied the open application programming interfaces (APIs) from eBay and Taobao e-commerce websites and captured real transaction data. Integrating multi-agent modeling, open APIs and social network analysis and propose a new way to study large-scale e-commerce systems.

The Web mining research is a convergingresearch area from several research communities, such as Databases, Information Retrieval and Artificial Intelligence. In this paper, Ratnakumar (2010) implement how Web mining techniques can be applied for the Customization i.e. Web personalization. Khosravi *et al* (2010) proposed an approach based on naïve Bayesian method for modeling and predicting users' navigation behavior. They used Web server logs as source data, for Web usage mining.

Shinde *et al* (2012) proposes a novel centering-bunching based clustering (CBBC) algorithm which is used for hybrid personalized recommender system (CBBCHPRS). Shen et al (2012) studies the problem of leveraging computationally intensive classification algorithms for large scale text categorization problems and proposed a hierarchical approach which decomposes the classification problem into a coarse level task and a fine level task. They demonstrate through extensive experimental evaluation that (1) the proposed hierarchical approach is superior to flat models, and (2) the data-driven extraction of latent groups works significantly better than the existing human-defined hierarchy.

## 3. WEB USAGE MINING USING PREPROCESING TECHNIQUE

The process of modifying the content and the arrangement of web site to the exact and particular needs of every user is Web Personalization.

Web Personalization process includes following steps:

a) Web data collection

b) Data categorization and modeling

c) Collected data analysis

d) Actions should be established.

Some similarity exists in Web page structure which belongs to same category. This general structure of web pages can be deduced from the placement of links, text and pictures (including videos and graphical records). This information can be easily pulled from an HTML text file.

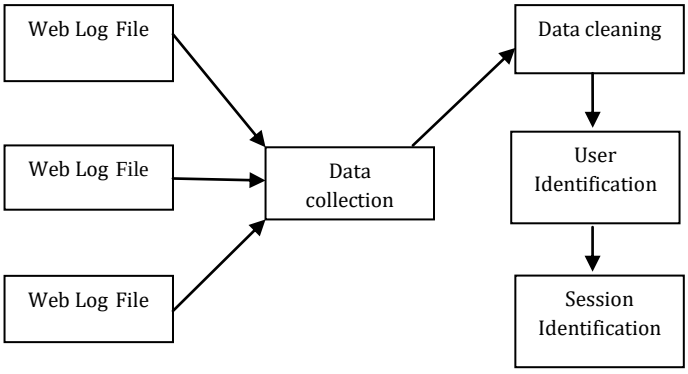
Information in the web site log is the main data source in the web usage mining and personalization process. In web log file, all the entries are recorded with date, time and the IP address, URI request, HTTP status code, etc.,. In terms of access patterns all the recorded details in the log data at web access reveal the navigational behavior knowledge of users.

Set of pages in a website is a collection of whole pages. During particular point, a single user clicks web page

is a user session. Practically, task can be contributed through set of visited pages which is exhibited and dominated by one specific navigational task. Depending upon the user visitor click number this navigational preference on individual page is given by its significant weight value. The user sessions (or called usage data), which are chiefly collected in the server logs, can be translated into a processed data format for the function of analysis via a data preparing and cleansing procedure. In one word, usage data is an accumulation of user sessions, which is in the form of weight distribution complete the page blank.

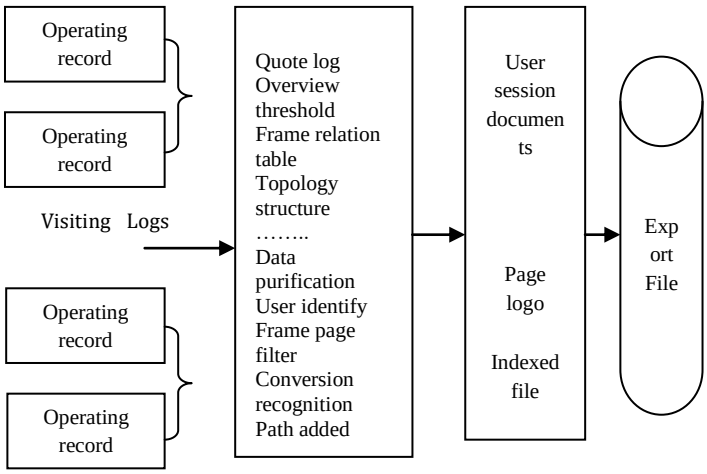
## Data Preprocessing

The data preparation process is often the most time consuming and computationally intensive step in the Web usage mining process. The process may involve preprocessing the original data, integrating data from multiple sources, and transforming the integrated data into a form suitable for input into specific data mining operations. This process is known as *data preparation.*




## 3.1 Improve the filtration method of Frame page

Applying the filtration method of Frame page can effectively eliminate the influence of Frame page on the log mining. However, the records of web log mining are very large. The filtering algorithm of Frame page in literature is judging whether each page of each user's session is Frame or SubFrame and deleting the sub-frame which has been judged one by one. Moreover, because of the deleting of SubFrame page, the upgraded site structure must be used in the following. Although interest degree has been added compared with the general preprocessing technology, the efficiency is still comparatively low and the spending also has been increased. Besides, SubFrame is deleted during the filtration, and it is also debatable that whether SubFrame can get full recovery in later path supplement. With consideration of decision Hierarchical algorithm has the property of allowing quick sort of multi-granularity layer, therefore, it can solve this problem better.




## 3.2 Particle Swarm based Hierarchical Web Session Clustering

In this proposed technique, modification of the session vector is done by adding new parameter user agent. This work used similarity measures such as *Spearman Distance* (SD) instead of Euclidean Distance (ED) alone.

$$D_{ij} = \sum_{k=1}^{n} (X_{ik} - X_{jk})^2 \quad (1)$$

In first step, session vector is to be first particle by initializing each particle. In next step, depending on the personal best positions (pBest) particles the iteration of the particles and nomination of the winning session takes place. The new positions and velocity of the particle is very important for applying Swarm. For finding the proper position, updating of velocity and position takes place in each iteration. In the final step, they apply adapted agglomerative algorithm to the winning session. This winning session gives the average linkage calculation between two clusters which is single input clusters to agglomerative as per Eq. 2. The proposed hierarchical sessionization algorithm based PSO and Agglomerative algorithm is explained.

$$Dist\,(C_i, C_j) = \frac{\sum X \in C_i \sum Y \in C_j \; Dist\,(X,Y)}{|C_i| \times |C_j|} \quad (2)$$

**Algorithm based particle swarm**

1. S= <s1,s2,s3,……,sn> where n <> 0
2. – particles as sessions
3. % Initializes positions of particle to session
4. Begin
5. For it = 1 in max_it
6. For i=1 in total session
7. For i=1 in max_part
8. pbest (I,j) = $\sum_{j=1}^{max\_part}(X_{ik} - X_{jk})^2$
% calculate Spearman Distance.
9. End j loop;
10. if min_dist <= pbest then
11. win_dist = min_dist
12. else
13. win_dist = pbest
14. end if
15. org_dist = $\sum_{i=1}^{total\_session}(X_i)^2$ % calculate the initial position of sessions
16. vf = vi + c1 * rand() * (pbest – xi)
17. xi = x + vf % update position of each session
18. End i loop
19. Copy win_dist sessions
20. End it loop
21. % Hierarchical clustering of wining Sessions
22. Given Set of wining particles (sessions)
23. Initialized cutoff = 0 – minimum distance to form clusters
24. Begin
25. For i=1 to total_clusters
26. For j=i+1 to total_clusters
27. Calculate the pair wise distance matrix
28. If distance < cutoff then
29. Merge_clusters (I,j)
30. End if
31. End j loop
32. End i loop
33. End
34. End

## 4. EXPERIMENTAL RESULT

The experimentation was done by using the proposed hierarchical algorithm on measures such as Euclidean Distance (Spearman Distance). For Euclidean base result evaluation, also performed the experiment. For reference of comparison, same parameters of session vectors are used. In a preprocessing step, Data Cleaning Algorithm was applied and the cleaning solutions are summed up in Table 1. Approximately, 20% were studied as irrelevant in entries of log and were transferred from the web logs.

**Table 1: Clean Data Records**

| Data | No of Log Records | Cleaned Log Records | % of Cleaned Log Records |
|---|---|---|---|
| 1 | 34000 | 5034 | 15 |
| 2 | 32000 | 7504 | 23 |
| 3 | 38000 | 6621 | 17 |
| 4 | 36000 | 8706 | 24 |
| Total | 140000 | 27919 | 20 |

For session identification they used the session vector (*Session ID; Session Length; Number of Pages Visited; Data Downloaded; User Agent*). They assumed the 30-minute timeout for sessions and took the number of pages visited > 10 and obtained a total of 625 unique sessions. Out of which 7 sessions have second or third episode. Average session length is 6 minutes, average number of pages visited in a session is 25, and average data transferred in a session is about 1.15 MB. Table 2 presents the details of session algorithm applied on cleaned and filtered web log

**Table 2: Summary of Session Data**

| | |
|---|---|
| Total No Session | 625 |
| Average Session Length | 5.90 minutes |
| Average Page visited in session | 25 |
| Average Data Transferred in session | 1.15 MB |

## 5. CONCLUSION

The main aim of the research was to improve the web log image and structured data in different stages of operation. They used *"Euclidean Distance" (Spearman Distance)* and PSO algorithm with agglomerative for achieving hierarchical sessionization of sessions. Therefore, the declaration concluded that for session clustering, there must be due weightage for proper similarity measures. For clustering phase, here applied the merger of PSO and agglomerative algorithm. In the first half of clustering algorithm, set of wining session is obtained by applying PSO algorithm. In the second half, agglomerative algorithm is applied for hierarchical sessionization of a web session.

## References

[1] Tyagi, Navin Kumar, A. K. Solanki, and Sanjay Tyagi. "An algorithmic approach to data preprocessing in web usage mining." *International Journal of Information Technology and Knowledge Management* 2, no. 2 (2010): 279-283.

[2] Thorleuchter, Dirk, and Dirk Van den Poel. "Using Webcrawling of Publicly Available Websites to Assess E-commerce Relationships." In *SRII Global Conference (SRII), 2012 Annual*, pp. 402-410. IEEE, 2012.

[3] Chitraa, V., Dr Davamani, and Antony Selvdoss. "A survey on preprocessing methods for web usage data." *arXiv preprint arXiv:1004.1257* (2010).

[4] Chou, Pao-Hua, Pi-Hsiang Li, Kuang-Ku Chen, and Menq-Jiun Wu. "Integrating web mining and neural network for personalized e-commerce automatic service." *Expert Systems with Applications* 37, no. 4 (2010): 2898-2910.

[5] Mei, Li, and Feng Cheng. "Overview of Web mining technology and its application in e-commerce." In *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, vol. 7, pp. V7-277. IEEE, 2010.

[6] Dimopoulos, Costantinos, Christos Makris, Yannis Panagis, Evangelos Theodoridis, and Athanasios Tsakalidis. "A web page usage prediction scheme using sequence indexing and clustering techniques." *Data & Knowledge Engineering* 69, no. 4 (2010): 371-382.

[7] Kularbphettong, Kobkul, Gareth Clayton, and Phayung Meesad. "A Hybrid System based on Multi-Agent System in the Data Preprocessing Stage." *arXiv preprint arXiv:1003.1792* (2010).

[8] Pamnani, Rajni, and Pramila Chawan. "Web Usage Mining: A research area in Web mining." *Proceedings of ISCET* (2010): 73-77.

[9] 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.

[10] Rao, VVR Maheswara, and Dr V. ValliKumari. "An Enhanced Pre-Processing Research Framework For Web Log Data Using A Learning Algorithm." *NeTCoM 2010, CSCP 01* (2010): 01-15.

[11] Piao, Chunhui, Xufang Han, and Harris Wu. "Research on e-commerce transaction networks using multi-agent modeling and open application programming interface." *Enterprise Information Systems* 4, no. 3 (2010): 329-353.

[12] Ratnakumar, A. Jebaraj. "An Implementation of Web Personalization Using Web Mining Techniques." *Journal of Theoretical and applied information technology* 18, no. 1 (2010): 67-73.

[13] Khosravi, Mahdi, and Mohammad J. Tarokh. "Dynamic mining of users interest navigation patterns using naive Bayesian method." In *Intelligent Computer Communication and Processing (ICCP), 2010 IEEE International Conference on*, pp. 119-122. IEEE, 2010.

[14] Shinde, Subhash K., and Uday Kulkarni. "Hybrid personalized recommender system using centering-bunching based clustering algorithm." *Expert Systems with Applications* 39, no. 1 (2012): 1381-1387.

[15] Shen, Dan, Jean-David Ruvini, and Badrul Sarwar. "Large-scale item categorization for e-commerce." In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 595-604. ACM, 2012