Enhanced User Navigation Prediction using Ensembling Clustering and Classification in weblog files

Dr.V. Sujatha[1] Dr. M.Punithavalli[2] and Dr.Ranjit Jeba Thangiah[3]

[1]Dean, Associate Professor:Department of Computer Applications
CMS College of Science and Commerce
Coimbatore,India
sujathapadmakumar4@gmail.com

[2]Associate Professor:Department of Computer Applications
Bharathiyar University
Coimbatore,India
mpunitha_srcw@yahoo.co.in

[3]Head, Associate Professor: Department of Computer Applications
Karunya University
Coimbatore,India
ranjit@karunya.edu

ABSTRACT

Web usage mining is the art of discovering navigation patterns of users from web log data. Next web page prediction, a task of web usage mining, is used to envisage future requirements of the user during surfing. This paper presents an ensemble prediction system, that uses ensemble clustering and ensemble classification. The proposed system uses heterogeneous clustering ensemble model to group similar browsing sequences together, which is then used by a heterogeneous classification ensemble model to predict future requests of the user. The goal of this combination process is to improve the quality of individual data clustering and classification. Experimental results demonstrate that the combination of ensemble is efficient in terms of prediction accuracy and can be used by web masters to attract users.

KEYWORDS

Web Mining, Weblogs, Classification, Clustering, Navigation Pattern & Prediction

1. INTRODUCTION

Web mining is an area of data mining, which employs techniques of machine learning algorithms on web (Internet) documents [1]. It is used to study various aspects of a website and recognize the relationships and patterns in user behavior in order to get an insight into crucial information. There are three types of web mining, namely, web content mining, web usage mining and web structure mining. This research is focused on web usage mining. Web Usage Mining is involved in mining usage characteristics of the users of web applications.

World Wide Web is the large source of online data, which includes text, images, videos, audio, etc., and is currently facing explosive growth of information. The information thus present in WWW lacks integrated structure or scheme, which poses serious drawbacks. From the user's perspective, it is very difficult to extract useful knowledge from the huge amount of information and secondly, it is also difficult to extract for the users to access relevant information efficiently. From business point of view [2], the webmasters and administrators find it difficult to organize the contents of the websites to cater to the needs of the users.

All these problems can be solved if the web navigation behaviour of an user can be understood. One way to extract such information from WWW is to perform navigation pattern analysis on web users' access data using data mining techniques. This knowledge can be extracted by analyzing the historical data, which are stored into files that are automatically created and maintained by servers. These files are called "Web log data", containing information about each and every hit made to a website. A hit information includes each view of a HTML page, image or other object. It typically stores information like IP address, user id and password, date and time stamp, status field indicating whether a request is successful or not, size of the file being transferred, referring URL and finally, the name and version of the browser being used. The extracted information can be used by many applications like personalization [3], website redesign or site improvement, business intelligence, navigation pattern behaviour prediction and web page recommender system.

2. PRE-PROCESSING

The pre-processing step of EN2PWD performs four major tasks. They are cleaning of web log data, path filling, user and session identification and clustering session. In general, most of the researches use only cleaning and user & session identification in web usage mining. In this research, two additional steps, namely, path filling and clustering user session are included. The inclusion of these two steps will improve the performance of the subsequent steps of EN2PWD [4].

2.1. Cleaning of Data

This task removes irrelevant entries in web log by removing all image, video and audio entries, removing all unsuccessful HTTP entry codes (Status code ≠ 200), retaining only those entries have GET and POST in request method field, removing all entries with blank IP addresses and remove all entries whose depth level is more than five.

2.2. User and Session Identification

User Identification is performed using IP address. A user session is a delimited set of pages visited by the same user within the duration of one particular visit to a website. Session identification is carried out using the assumption that if in a certain predefined period of time between two accesses is exceeded, a new session starts at that point. Identify unique users in a session who have in the website for more than 30 minutes. Each session is assumed to be of duration 30 minutes and a new session starts after each 30 minutes.

2.3. Path Filling

Another potentially important pre-processing task after sessionization is path completion. Path completion is process of adding the page accesses that are not in the web log but those which be actually occurred. The proposed path filling method performs path completion using an optimized 2-Way hash structure. This hash structure is used to represent user accessed page sequence using an Access History List (AHL). The 2-way has structure is optimized to fill the path of only those users who are actually interested in the website. These users are termed as potential users.

The final step of preprocessing is the clustering of session using rough set clustering technique. The advantage of clustering the sessions is that the data mining algorithms can now be applied on the most probable user sessions and thus prediction process [5] can provide appropriate set of pages that can help systems that use next page access information. This also reduces the time complexity of the overall system.

Identification of potential users [6] reduces the time spent on analyzing irrelevant entries and thus improves the performance of EN2PWD. For this purpose, this research proposes a potential user identifier that combines a rule based algorithm with an interest measure.

3. Prediction System

The prediction system is composed of two steps, namely, clustering and classification. The research work proposes three categories of clustering based classification systems and two heterogeneous ensemble based classification system. For this purpose, three clustering algorithms (Ant based clustering algorithm, Improved Pairwise Nearest Neighbour algorithm and Graph partitioning algorithm) [8] and three classification algorithms (Maximum Likelihood Classification Algorithm, Longest Common Sequence Classification Algorithm and Markov Model based Classification Algorithm) are used. The same algorithms are used during the design of the respective ensemble systems. The aggregation method used by ensemble systems [7] is the majority voting algorithm. The proposed clustering based classification systems are

(i) Ant-based Models
  a. Ant-based with MLC (AMLC)
  b. Ant-based with LCS (ALCS)
  c. Ant-based with MM (AMM)
  d. Ant-based with ECLA (AECLA)

(ii) Graph Partition-based (GP) Models
  a. GP with MLC (GPMLC)
  b. GP with LCS (GPLCS)
  c. GP with MM (GPMM)
  d. GP with ECLA (GPECLA)

(iii) Improved PNN (IP) Models
  a. IP with MLC (IPMLC)
  b. IP with LCS (IPLCS)
  c. IP with MM (IPMM)
  d. IP with ECLA (IPECLA)

The proposed heterogeneous clustering [9] based classification is designed using the two ensemble system build with the three clustering and classification algorithms and is termed as ECLU-ECLA system.

(i) Ensemble clustering approach using majority voting algorithm (ECLU)
(ii) Ensemble classification approach using majority voting algorithm (ECLA)

4. Experimental Results

Various experiments were conducted to analyze the efficiency of each of the steps in EN2PWD. For this purpose, web logs from www.microsoft.com were collected. It records 37,711 randomly selected anonymous users of the site of which 32,711 are given as training set and 5000 as test set. For each user, the data lists all the areas of the web site that user visited in a one-week timeframe.

Table 1 shows the effect of pre-processing on number of transactions and memory usage while using three different sized log data.

Table 1. Effect of Pre-processing

	1 Day		3 Days		1 Week	
	Before	After	Before	After	Before	After
No. of Transactions	7000	1420	11987	4320	34000	11381
Memory Used (MB)	1.76	0.48	1.95	0.61	2.47	1.12

From the table, it is evident that pre-processing log file reduces both the memory used to store the log file and the number of transactions in a tremendous fashion. Even when supplied with a log data set having 34000 entries (1 week data), a 66.53% reduction was envisaged after pre-processing. Similarly, while considering the storage space saved, the pre-processed web log data showed 54.66% gain when compared to raw log data. This shows that both time and memory requirement can be saved while using the pre-processing step in EN2PWD.The effect of identifying potential users to prune non-potential users is shown in Figure 1.

From the figure, after detecting the potential users through potential user identification, the number of entries in the transaction database decreased. It could also be seen that more than 70 per cent of the users are pruned out after classification, thus reducing the size of the web log data file. The next stage of experiments focuses on analyzing the performance of the proposed next web page prediction algorithms. Four performance metrics, namely, accuracy, coverage, F-Measure and speed are used for this purpose.

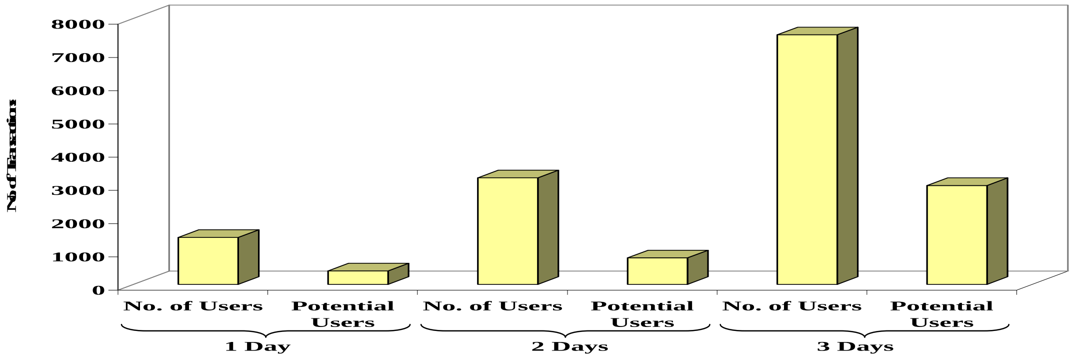

Figure 1. Effect of Pruning

The navigation patterns are identified using the clustering algorithm and the grouped patterns are then divided into two sets. The first set is used for generating prediction and the second set is used to evaluate the predictions. Let as$_{np}$ denote the navigation pattern obtained for the active session‘s’ and let T be a threshold value. The prediction set is denoted as P(as$_{np}$, T) and the evaluation set is denoted as eval$_{np}$ The three parameters can then be calculated using Equations (1), (2) and (3).

$$\text{Accuracy} = \frac{|\,P(as_{np},T) \cap eval_{np}\,|}{|\,P(as_{np},T)\,|} \quad (1)$$

$$\text{Coverage (P(asnp, T))} = \frac{|\,P(as_{np},T) \cap eval_{np}\,|}{|\,eval_{np}\,|} \quad (2)$$

$$\text{F1(P(asnp, T))} = \frac{2x\text{Accuracy}(P(asnp,T))\ \ x\ \ \text{Coverage}(P(asnp,T))}{\text{Accuracy}(P(asnp,T)) + \text{Coverage}(P(asnp,T))} \quad (3)$$

Accuracy measures the degree to which the prediction algorithms produce accurate recommendations while coverage measures the ability of the prediction algorithms to produce all of the pageviews that are likely to be visited by the user. The F1 measure attains its maximum value when both accuracy and coverage are maximized. Speed of prediction is the execution time taken by the algorithms to predict the next page. Tables 2-4 presents the performance of the clustering based prediction algorithms with respect to the four selected performance metrics. The results are grouped according to the prediction algorithm used.

Table 2. Performance of Clustering Based MLC Prediction Algorithms

Algorithms	Accuracy	Coverage	F1 Measure	Speed
MLC	70.78	0.8105	1.603	10.41
AMLC	77.65	0.7532	1.492	9.31
IPMLC	73.68	0.6277	1.245	9.97
GPMLC	79.71	0.5418	1.076	9.26

Table 3: Performance of Clustering Based LCS Prediction Algorithms

Algorithms	Accuracy	Coverage	F1 Measure	Speed
LCS	76.71	0.7762	1.537	7.41
ALCS	81.63	0.7073	1.402	5.54
IPLCS	78.74	0.5689	1.13	5.62
GPLCS	82.67	0.4677	0.93	5.58

Table 4: Performance of Clustering based MM Prediction Algorithms

Algorithms	Accuracy	Coverage	F1 Measure	Speed
MM	74.94	0.7426	1.471	8.26
AMM	79.67	0.7164	1.42	6.56
IPMM	75.71	0.6003	1.191	6.71
GPMM	81.66	0.5092	1.012	6.34

From Table 2 results, it can be seen that the while all the proposed models perform better than the existing maximum likelihood classification algorithm, the GPMLC prediction algorithm produces maximum performance in terms of all the four selected performance metrics. The GPMLC model shows a performance gain of 11.20% with respect to accuracy when compared to the existing MLC algorithm.

Table 3 data reveals that again all the proposed prediction algorithms show improved results when compared to the existing LCS algorithm. However, GPLCS algorithm combining graph partitioning clustering algorithm and LCS algorithm is more successful when compared to other three LCS based clustering-classification algorithms. The GPLCS algorithm showed an efficiency gain of 7.21% over LCS algorithm while considering accuracy performance parameter.

Table 4 tabulates the results obtained by the markov model prediction algorithms while using the three selected clustering algorithms, namely, ant-based clustering, improved PNN based clustering and graph partition based clustering algorithms. The results reveal that performance of prediction has improved when markov model is combined with clustering algorithm and the prediction algorithm that combines markov model with graph partitioning clustering algorithm shows high performance when compared to other proposed algorithms. The GPMM algorithm shows an accuracy efficiency gain of 8.23% when compared with MM prediction algorithm.

While comparing the winning algorithms in each category, the performance of GPLCS algorithm is high when compared to GPMCL and GPMM prediction algorithms. The GPLCS algorithm showed prediction accuracy efficiency gain of 3.58% when compared to GPMCL and 1.22% when compared with GPMM algorithm.

Figures 2a-2d shows the performance of the clustering based ensemble classification algorithms with respect to the four selected performance metrics, namely, accuracy, coverage, F1 measure and speed of prediction. From the results, it is evident that the ensemble system combining MLC, LCS and MM prediction algorithms with graph partitioning clustering algorithm produces best results in terms of all selected performance metrics. The GPELCA prediction algorithm showed an efficiency gain of 2.18% when compared with AECLA and 8.86% when compared with IPECLA prediction algorithms while considering accuracy parameter.

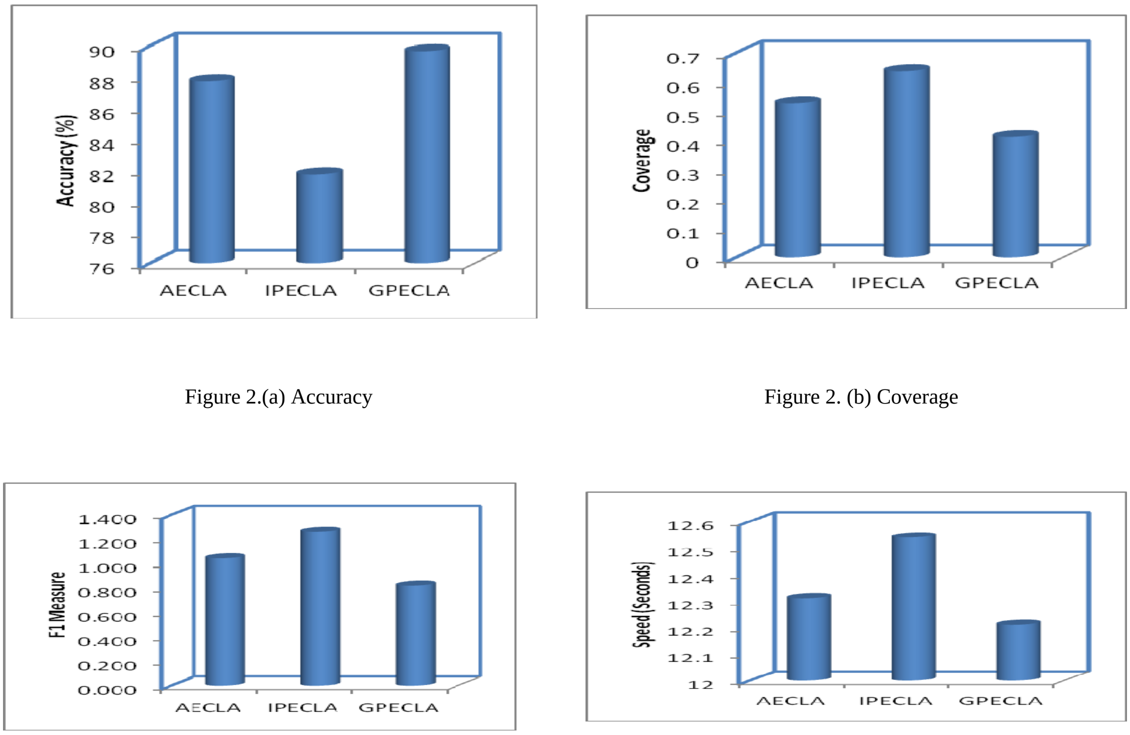

Figure 2.(a) Accuracy

Figure 2. (b) Coverage

Figure 2.(c) F1 Measure

Figure2.(d) Speed

Figure 2. Comparison of Clustering Based Ensemble Prediction Algorithms

Table 5. Performance of Ensemble Clustering Based Ensemble Prediction Algorithm

Algorithm	Accuracy	Coverage	F1 Measure	Speed
ECLUECLA	94.64	0.3503	0.69801637	17.37

The results show that the performance of ECLUELA prediction algorithm is very high and is best suited for predicting next page. On average, the CLU-CLA method showed an efficiency gain of 1.96% when compared with ECLULCS prediction algorithm and 5.27% when compared with GPECLA prediction algorithm. Thus, it can be concluded that the goals of the research are met by the proposed EN2PWD using preprocessing, pruning and CLU-CLA hybrid ensemble model.

5. Conclusion

This research work is focused on designing and developing next web prediction algorithms to improve the browsing experience of users. The proposed algorithm was built in three steps, namely, preprocessing, clustering and classification. The preprocessing step transforms the raw web log data into a form that can be directly used by the clustering and classification (prediction) algorithms. The clustering step is used to improve the performance of classification algorithm in terms of accuracy and time. The applicability of three clustering algorithms, namely, ant based clustering, improved pairwise nearest neighbour algorithm and graph partitioning algorithm, were studied for grouping web log data. Similarly, three classification algorithms (Maximum Likelihood Classification Algorithm, Longest Common Sequence Classification Algorithm and Markov Model based Classification Algorithm) were analyzed for predicting next web page. Using these clustering and classification algorithms, ensemble clustering and ensemble classification systems were designed which were combined to form ensemble clustering based classification systems. Experimental results prove that the proposed ensemble model is successful in next web page prediction in terms of accuracy, coverage and F1 measure.

References

[1] Yan, Z., Zhang, P. and Vasilakos, A.V. (2014) A survey on trust management for Internet of Things, Journal of Network and Computer Applications, Vol. 42, pp. 120-134.

[2] Ryals, L. and Knox, S. (2001) Cross-functional issues in the implementations of relationship marketing through CRM, European Management Journal, Vol. 19, Issue 5, pp. 534-542.

[3] Pagar, Y.S., Mote, V.R. and Bramhane, R.S. (2012) Web Personalization using Web Mining Techniques, IJCA Proceedings on Emerging Trends in Computer Science and Information Technology, Vol. 1, pp. 1-4.

[4] Sujatha, V. and Punithavalli, M. (2012) Improved user navigation pattern prediction technique from web log data, International Conference on Communication Technology and System Design, Procedia Engineering, Vol. 30, pp. 92-99

[5] Anitha. A. A New Web Usage Mining Approach for Next Page Access Prediction, International Journal of Computer Applications, Vol. 8, No. 11, 2010, pp. 7–10.

[6] Langhnoja, S.G., Barot, M.P. and Mehta, D.B. (2013) Web Usage Mining to Discover Visitor Group with Common Behavior Using DBSCAN Clustering Algorithm, International Journal of Engineering and Innovative Technology, Vol. 2, Issue 7, pp. 169-173.

[7] Hu, X. and Yoo, I. (2004). Cluster Ensemble and Its Applications in Gene Expression Analysis. In Proc. Second Asia-Pacific Bioinformatics Conference (APBC2004), Dunedin, New Zealand.CRPIT, 29. Chen, Y.-P. P., Ed. ACS. 297-302.

[8] Jalali, M., Mustapha, M., Mamat, A. and Sulaiman, M.N.B. (2008) A new clustering approach based on graph partitioning for navigation patterns mining, 9th International Conference on Pattern Recognition, pp. 1-4.

[9] Hadjitodorov, S., Kuncheva, L. I and Todorova, L.P. (2006) Moderate Diversity for Better Cluster Ensembles. Information Fusion Journal, pp.264275.