ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Analyzing the User Navigation Pattern from Weblogs

P.G. Om Prakash *, A.Jaya
* Research Scholar CSE Department,* Professor CA Department, B.S.A.University, Chennai.

Abstract
In a real world lot of users attracted towards online purchasing, so lots of transactions are going on in the websites. A weblog contains series of entries updated frequently by the user while accessing the website. A Weblog comprises of various entries like IP address, Status Code, number of bytes transferred and timestamp etc. Based on user interest related and unrelated data can be classified. The related data can be considered as success response, while the unrelated data can be considered as failure response. This research work aims to analyze the pattern of user navigation while browsing, for that web usage mining must be analyzed. The process of Web Usage Mining consisting steps: Data Collection, Pre-Processing, Pattern Discovery and Pattern Analysis to get user navigation pattern that will help us to predict the user behavior and it reduces the mining time.
Keywords: User navigation, web mining, user behaviour, traversal pattern, prediction accuracy, Data Mining.

Introduction
In current situation explosive growth of knowledge available on internet makes the users to access the information day by day. It becomes much more difficult for users to access relevant information efficiently. Analyzing and modelling web navigation behaviour is helpful in understanding demands for online users. Web mining is the application of data mining technique to extract and analyze useful information from web data [2]. Based on the kind of data, web mining can be classified in to three different categories namely web content mining, web structure mining and web usage mining. web content mining is the discovery of useful information from the contents of web documents such as image, text, audio, video etc. Web structure mining focus on analyzing the physical link structure of websites. Web usage mining analyzes the browsing activity.

The web usage data consists of data from weblog. The user accessing information from the websites are stored as logs. The log contains series of user transactions which are frequently updated whenever the user accesses the website [3]. The prediction of user behaviour can be identified only through logs. The weblog contains unstructured format, so convert to raw weblog to processed weblog using data preprocessing[4], the data preprocessing includes Data Cleaning, User identification, Session Identification, content retrieval and path completion to get user navigation pattern.

Web usage mining prediction process is structured according to online and offline with respect to web server activity [5]. Offline components built the knowledge base by analyzing historical data, such as server access log file or weblogs which are captured from server. Weblog used in online component for capturing the intuition list, whenever that user comes online for next time.

The remaining part of the paper is organized as follows, Section 2.0 is deals related work, Section 3.0 describes architecture diagram and Section 4.0 gives the experimental results. Finally Section 5.0 concludes the paper by giving to future directions of research in this area.

Related work
Web usage mining systems proposed to predict the user navigation behavior and the preferences in website, User navigation system uses data preprocessing Dr.A.R.Patel and Renata Ivancsy [1], The weblog contains unstructured format, so convert to raw weblog to processed weblog using data preprocessing, the data preprocessing contains Data

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Cleaning, User identification, Session Identification, content retrival and path completion to get user navigation pattern. After getting the processed log, the given log is converted in to web traversal pattern. Hong Cheng and Xifeng Yan [2] uses classification algorithm to classify the processed log into Frequent Sequence, Semi-frequent Sequence and In-frequent Sequence. D.Kerana Hanirex and Dr. M.A. Dorai Rangaswamy [4] finds the frequent item sets by partitioning the database transactions into clusters. Clusters are formed based on the similarity measures between the transactions.Then it finds the frequent itemsets with the transactions in the clusters directly using the improved clustering algorithm which further reduces the number of scans in the database and hence improve the efficiency.

Mobasher [11] present Web personalizer a system which provides dynamic recommendations, as a list of hypertext links, to users. The analysis is based on anonymous usage data combined with the structure formed by hyperlinks of the site. Data mining techniques i.e. clustering, sequence pattern discovery and association rules are used in preprocessing phase in order to obtain aggregate usage profiles. In this phase Web server logs are converted into clusters of visited pages, and cluster made up of set of pages with common usage characteristics. The online phase considers active user session in order to find matches among user's activities and discovered usage profiles. Matching entries are used to compute a set of recommendations which will be inserted into last requested page as list of hypertext links. Web Personalizer is a good example of two tier architecture for Personalization Systems. Baraglia and Palmerini proposed a WUM system called SUGGEST, that provide useful information to make easier the web user navigation and to optimize the web server performance. SUGGEST adopts a two level architecture composed of offline creation of historical knowledge and online engine that understands user's behavior.
Joachims, Juhne [26] and adaptive web site agents Pazzani and Billsus [28] are examples of web tour guides, agents that help visitors browse a site by suggesting which link each visitor should view next. With the assistance of a tour guide, visitors can follow trails frequently viewed by others and avoid becoming lost. However, tour guides assume that every page along the trail is important, and typically are limited to only suggesting which link on a page to follow next as opposed to creating shortcuts between pages.

Fu and Perkowitz and Etzioni [29] suggest pages to visit based on page requests co-occurrent in past sessions. These algorithms suggest the top m pages that are most likely to co-occur with the visi-tor's current session, either by presenting a list of links Sur-fLen or by constructing a new index page containing the links PageGather. However, both of these systems assume the visitor can easily navigate a lengthy list of shortcuts, and thus provide perhaps dozens of suggested links. MINPATH improves on these algorithms by factoring in the relative benefit of each shortcut, and suggesting only the few best links specific to each page request.

The predictive web usage models we present are related to previous works on sequence prediction and web usage mining. These works are too numerous to review here, but we mention two closely related ones. Most similar to our own work, Cadez, 2000 [30] is a system for visualizing clusters of web visitors using a mixture of Markov models. We apply similar models to web behavior, although our goal is to build predictive structures, while Web-CANVAS emphasizes visualizing the clusters themselves. Sarukkai uses a Markov model of web usage to suggest the most probable links a visitor may follow, and notes the need to reduce the size of the model by clustering the URLs. Our work explores this model as well as many others, and uses the expected savings of a link, not just the link probability, to sort the resulting suggestions.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Architecture diagram

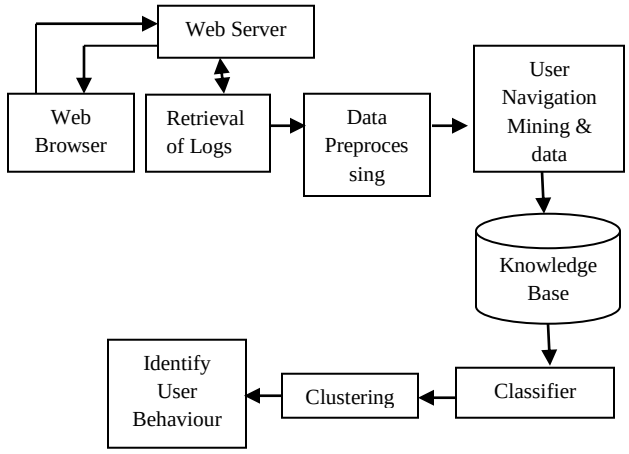

Figure 1 Conceptual design & Analysis of User Pattern

The user accessing the webserver from that the logs will be generated. The log is called as unstructured log. The raw log is converted into web traversal pattern by data preprocessing, the web traversal pattern is called as knowledge base, it contains previous user navigation path sequence. The classification algorithm is get from the knowledge base it converts the logs in to frequent sequence, semi frequent sequence and infrequent sequence. The clustering phase is to grouping the frequent sequence and semifrequent sequence, the NaiveBayes is to utilizes the previous results and reduces the number of paths and improves prediction accuracy.

Classifier

The classifier utilizes IncSpan algorithm, the given processed log is converted in to Frequent Sequence, Semi-frequent Sequence and In-frequent Sequence[9].

D – original database
D' – Appended database
min_sup – threshold
μ - buffer ratio
FS – set of frequent sequences
SFS – set of semi-frequent sequences
>min_sup – sequence is frequent
μ <= 1 – sequence is semi-frequent
< μ * min_sup – sequence is in frequent

Input: An appended database D', min_sup, μ, frequent sequences FS in D, semi-frequent sequences SFS in D.
Output: FS' and SFS'

1. FS' = Ø, SFS' = Ø
2. Scan LDB for single items;
3. Add new frequent item into FS';
4. Add new semi-frequent item into SFS';
5. for each new item I in FS' do
6. PrefixSpan(I,D'|i , μ * min_sup, FS', SFS');
7. for every pattern p in FS or SFS do
8. check sup(p);
9. if sup(p) = sup(p) + sup(p) >= min_sup
10. insert(FS', p);
11. if sup(p) >= (1-μ)min_sup
12. PrefixSpan(p,D'|p, μ * min_sup, FS', SFS');
13. Else
14. insert (SFS', p);
15. return;

The most common form of pattern analysis is combining WUM tools with a knowledge query mechanism such as SQL. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns

Clustering

The Clustering model is used as PAFI algorithm, the partition algorithm frequent itemset is to cluster the Frequent Sequence, Semi-frequent Sequence to form Clustering.

Partition Algorithm for Mining Frequent Item sets (PAFI) using clustering technique.

This algorithm finds the frequent item sets by partitioning the database transactions into clusters. Clusters are formed based on the similarity measures between the transactions.

Then it finds the frequent itemsets with the transactions in the clusters directly using the improved Apriori algorithm which further reduces the number of scans in the database and hence improve the efficiency.

Apriori algorithm uses PAFI algorithm which further reduces the number of scans in the database and hence improve the efficiency.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Step 1Begin
Number of clusters(NOC)=count of transactions(COT)/N //N is random natural number
FOR i= 1 to NOC DO BEGIN
FOR each cluster Ci DO BEGIN
FOR each transaction tÎD DO BEGIN
Find t such that t having highest number of items
Put t in Ci
END
END
Return Clusters with 1 itemset.

Weblogs

The web log contains the details of previous user navigation, the samle log contains the details of [21]

(1) the user's IP address
(2) the remote logname of the user
(3) the access date and time
(4) the request method
(5) the URL of the page
(6) the protocol (HTTP 1.0, HTTP 1.1,etc.)
(7) the return code
(8) the number of bytes transmitted

in24.inetnebr.com - - [01/Aug/1995:00:00:01 -0400] "GET/shuttle/missions/sts-68/news/sts-68-mcc-05.txt HTTP/1.0" 200 410

Compiler that Transforms Weblog Data into Sessions

The first step for preparing the data is the transformation of the set of logs into sessions. We have defined a compiler that transforms a set of log entries L,

L = _L1, ...,L|L|_
Li = (IPi,LOGNAMEi, TIMEi,METHODi, URLi, PROTi,CODEi,BY TESi), ∀i/i = 1...|L|, (1)
into a set of sessions S,
S = _S1, ..., S|S|_ , (2)

where |L| is the number of log entries in L and |S| is the number of sessions of S. Each session is defined as a tuple (USER,PAGES):
Si = (USERi, PAGESi), PAGESi = {urli,1, ..., urli,pi} , i = 1... |S| , (3)
where *USER* identifies the user of the session, *PAGES* the set of pages requested,and *pi* is the number of pages requested by user *USERi* in session *Si*.

L – set of log entries
|L| - num of log entries in L
S – set of sessions
|S| - num of sessions in S

Input : L, Δt, |L|
Output : S, |S|
function Compiler(L, Δt, |L|)
for each Li of L
if METHODi is GET and URLi is WEB PAGE then
if ∃ Sk ∈ OPEN SESSIONS with USERk = USERi
if (TIMEi − END TIME(Sk)) < Δt then
Sk = (USERk, PAGESk ∪ URLi)
else
CLOSE SESSION (Sk)
OPEN NEW SESSION(USERi, (URLi))
end if
else
OPEN NEW SESSION(USERi, (URLi))
end if
end if
end for

Experiments

The proposed NaiveBayes algorithm was implemented using Weka. It is implemented with knowledge base which has 846 instances in log. The maximal compactness is 119 and the minimum compactness is 73. In total, 19 attributes were generated. These new attributes were used to adjust the entailment decision threshold and to evaluate the final system performance. The accuracy was calculated using the ratio between the number of Instances generated by the system and the total number of attributes. Fig.4.1 & 4.2 shows the sample output screen for Search Result.

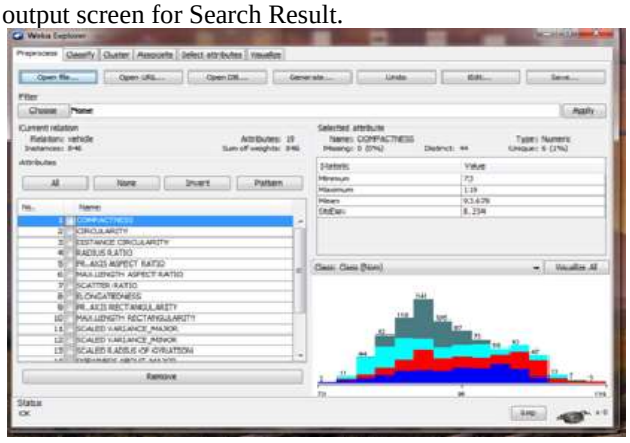

Figure 2 Compactness of vehicle

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

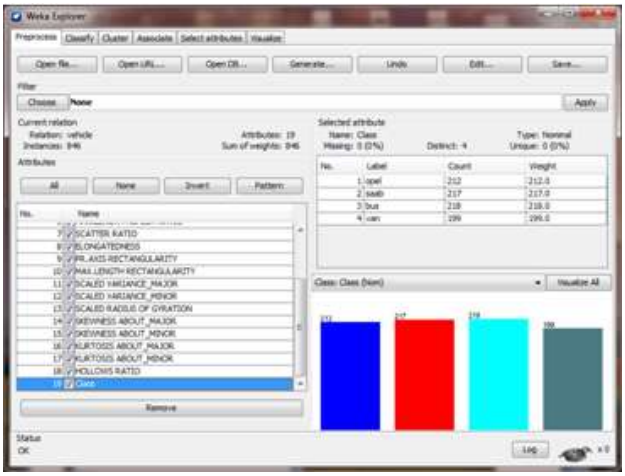

Figure 3 Class of vehicle

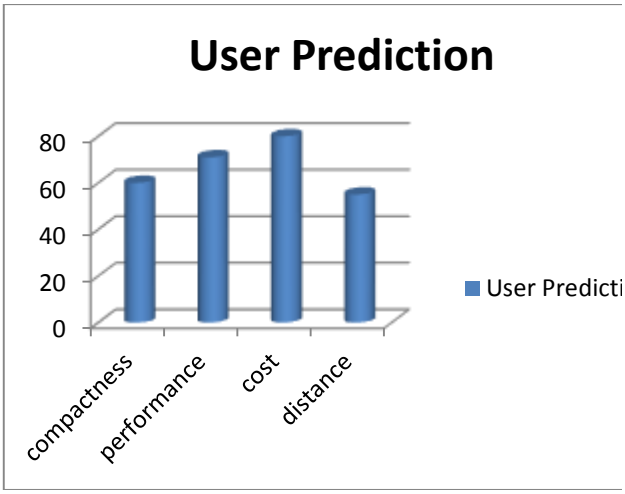

Figure 4 Intuition of car by user interest

The figure 4. shows the intuition of car by user interest. By analyzing, various attributes of car were generated. The major four attributes are calculated by ratio between the number of instance generated by the system and the total number of attributes. The user prediction can be generated using the majority of attributes such as compactness, performance, cost and distance. By using user prediction analysis, the business process can be improved.

Conclusion

Web service systems model is low-cost. The major drawback of the system is difficult to handle user navigation. Using NaiveBayes improves prediction accuracy, it utilizes the previous results and reduces the number of path from the weblogs & IncSpan algorithm is to reduce the mining time and utilizes the previous mining results.

References

1. *Ketul B. Patel, Dr.A.R. Patel ,"Process of web usage mining to find interesting patterns from web usage data,"www.ijctonline.comVol. 3, No. 1,Aug 2012.*
2. *Hong Cheng, Xifeng Yan, Jiawei HanIncSpan: Incremental Mining of Sequential Patterns in Large Database*
3. *pawel Weichbroth, Mieczyslaw Owoc, Michal PlesZkun "Web User Navigation Patterns Discovery from WWW server Log Files", IEEE 2012,*
4. *D.Kerana Hanirex, Dr. M.A. Dorai Rangaswamy" Efficient Algorithm for Mining Frequent Itemsets Using Clustering Technique," International Journal on Computer Science and Engineering, vol. 3, no. 3, March 2011.*
5. *X. Fang and C. Holsapple, "An Empirical Study of WebSite Navigation Structures' Impacts on Web Site Usability,"Decision Support Systems, vol. 43, no. 2, pp. 476-491, 2007.*
6. *J. Lazar, Web Usability: A User-Centered Design Approach. Addison Wesley, 2006.*
7. *D.F. Galletta, R. Henry, S. McCoy, and P. Polak, "When the Wait Isn't So Bad: The Interacting Effects of Website Delay, Familiarity, and Breadth," Information Systems Research, vol. 17,no.1,pp.20- 37, 2006.*
8. *J. Palmer, "Web Site Usability, Design, and Performance Metrics,"Information Systems Research, vol. 13, no. 2, pp. 151-167, 2002.*

   [9] V. McKinney, K. Yoon, and F. Zahedi, "The Measurement of Web- Customer Satisfaction: An Expectation and Disconfirmation Approach," Information Systems Research, vol. 13, no. 3, pp. 296-315, 2002.
9. *T. Nakayama, H. Kato, and Y. Yamane, "Discovering the Gapbetween Web Site Designers' Expectations and Users' Behavior,"Computer Networks, vol. 33, pp. 811-822, 2000.*
10. *M. Perkowitz and O. Etzioni, "Towards Adaptive WebSites:Conceptual Framework and*

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Case Study," Artificial Intelligence,vol. 118, pp. 245-275, 2000.

11. *J. Lazar, User-Centered Web Development. Jones and Bartlett Publishers, 2001.*
12. *Y. Yang, Y. Cao, Z. Nie, J. Zhou, and J. Wen, "Closing the Loop in Webpage Understanding," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 5, pp. 639-650, May 2010.*
13. *J. Hou and Y. Zhang, "Effectively Finding Relevant Web Pagesfrom Linkage Information," IEEE Trans. Knowledge and Data Eng.,vol. 15, no. 4, pp. 940-951, July/Aug. 2003.*
14. *H. Kao, J. Ho, and M. Chen, "WISDOM:Web IntrapageInformative Structure Mining Based on Document Object Model,"IEEE Trans. Knowledge and Data Eng., vol. 17, no. 5, pp. 614-627, May 2005.*
15. *B. Mobasher, R. Cooley, and J. Srivastava, "Automatic personalization based on Web usage mining"Communications of the ACM, vol. 43, pp. 142-151, 2000.*
16. *C. R. Anderson, P. Domingos, and D. S.Weld, "Adaptive Web Navigation for Wireless Device" Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, pp. 879–884, 2001.*
17. *I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "Visualization of navigation patterns on a Web site using model-based clustering," Proceedings of the sixth ACM SIGKDD internationalconference on Knowledge discovery and data mining, pp.280-284, 2000.*
18. *Dr.R.Lakshmipathy,V.Mohanraj, J.Senthilkumar, Y.Suresh, "Capturing Intuition of Online Users using a Web Usage Mining" Proceedings of 2009 IEEE International Advance Computing Conference (IACC 2009)Patiala, India, 6-7 March 2009.*
19. *J.Ben Schafer, Joseph Konstan, John Riedl "Recommender Systems in E-Commerce" GroupLens*
20. *Research Project Department of Computer Science and Engineering,University of Minnesota.*
21. *M. Perkowitz and O. Etzioni, "Towards adaptive Web sites: Conceptual framework and case study," Artificial Intelligence, vol. 118, pp. 245-275, 2000.*
22. *M. Jalali, N. Mustapha, A. Mamat, Md N. Sulaiman, "OPWUMP An architecture for online predicting in WUM-based personalization system", In 13th International CSI Computer Science, Springer Verlag,2008. 307*
23. *R. Baraglia and F. Silvestri, "Dynamic personalization of web sites without user intervention," Communications of the ACM,vol. 50, pp. 63-67, 2007.*
24. *I. V. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of navigation patterns on a web site using model based clustering. In Proc. 6th Intl. Conf. on Knowl-edge Discovery and Data Mining, 2000.*
25. *R. R. Sarukkai. Link prediction and path analysis using Markov chains. In Proc. 9th Intl. WWW Conf., 2000.*
26. *T. Joachims, D. Freitag, and T. Mitchell. WebWatcher: A tour guide for the World Wide Web. In Proc. 15th Intl Joint Conf. on Art. Int., 1997*
27. *J. Juhne,¨ A. T. Jensen, and K. Grønbæk. Ari-adne: a Java-based guided tour system for the World Wide Web. In Proc. 7th Intl. WWW Conf., 1998.*
28. *M.J. Pazzani and D. Billsus. Adaptive web site agentsIn Proc. 3rd Intl. Conf. on Autonomous Agents, 1999.*
29. *X. Fu, J. Budzik, and K. J. Hammond. Mining navigation history for recommendation. In Proc. 2000 Conf. on Intelligent User Interfaces, 2000.*
30. *M. Perkowitz and O. Etzioni. Adap-tive web sites: an AI challenge. In Proc. 15th Intl. Joint Conf. on Art. Int., 1997.*