© 2017 IJSRST | Volume 3 | Issue 1 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Enhancement in Next Web Page Recommendation with the help of Multi-Attribute Weight Prophecy

Prof. Umesh A. Patil, Avinash Kunnure, Vaibhav Herwade, Vijaykumar Dawarpatil, Satyawan Magar

Computer Science and Engineering, D. Y. Patil Technical Campus, Faculty of Engineering & Faculty of Management, Talsande, Kolhapur, Maharashtra, India

ABSTRACT

In Today internet world has increasing number of websites so it's the big task to get accurate data from large numbers of website. The web data mining is one of the challenging task .While performing the web page prediction pre-processing of the data from a web site. The necessity for predicting the user's needs in order to enhance the usability and user maintenance of a web site is more than marked now a day's lacking proper guidance, a visitor often wanders aimlessly without visiting significant pages, loses attention, and leaves the site earlier than expected. When they access the network, a large amount of data is generated and is stored in Web log files which can be used efficiently as many times user frequently searched the same type of Web pages recorded in the log files. These sequence can be considered as a web access pattern, valuable to find the user behavior Through this custom-made information, it's quite easy to forecast the next set of pages user might visit based on the previously searched patterns, thereby reducing the browsing time of a user.

Keywords: Web Usages Mining, Recommendation, Web Log Analysis, Session Based Predication, K-NN Algorithm

I. INTRODUCTION

Web page prediction is the web usage mining by performing pre-processing of the data from a web site. Web prediction is a classification problem which attempts to predict the most likely web pages that a user may visit depending on the information of the previously visited web pages. The need for prophecy the user's needs in order to enhance the usability and user maintenance of a web site. Web usage mining is widely used to discover the usage patterns from web log files. It deals with web log data which are taken from web servers, proxy server or clients cache.

The proposed web recommendation system is concept by which the previous or historical user navigation data is analyzed and based on the navigation technique;he next web page access is predicted. The proposed recommendation system has some relevant concepts such as behaviour analysis of user access patterns, personalization of data and predictive modelling. The behaviour of users accessed data is extracted using the K-mean clustering algorithm. Then search the similar user behaviours from the web log using KNN algorithm which analyse data in distance based function and most nearest patterns are listed with the help of user frequent patterns. From nearest frequent pattern, the time based data clustering is also prepared to amount of time spent on a particular URL in the entire log file. After evaluation of these parameters namely user navigational frequency and time based frequency a combine weight for all the URLs are evaluated. These weights are further sorted and by the rank of weights the next most possible web page is predicted.

This project is contributed that improves the recommendation accuracy based on the session of user's web access. It provides more appropriate recommended web page to the active user.

II. METHODS AND MATERIAL

1. Literature Review

Web Usage Mining is an important application of data mining technique used to discover interesting usage patterns from the Web logs to understand and serve the requirements of Web-based applications. A Web log

along with the identity of the user captures their browsing behaviour on a web site. Web prediction is a classification problem which attempts to predict the most likely web pages that a user may visit depending on the information of the previously visited web pages. The K-Nearest-Neighbor (KNN) classification method has been trained to be used on-line and in Real-Time to identify clients/visitors click stream data, matching it to a particular user group and recommend a tailored browsing option that want to the specific user at a particular time.

2. Proposed Work

This project focus is to develop a next web page recommendation system using web access logs. The data in the log files of the server about the actions of the users can not be used for mining purposes in the form as it is stored. For this reason the data should be pre-processed to improve the efficiency and ease of the mining process.

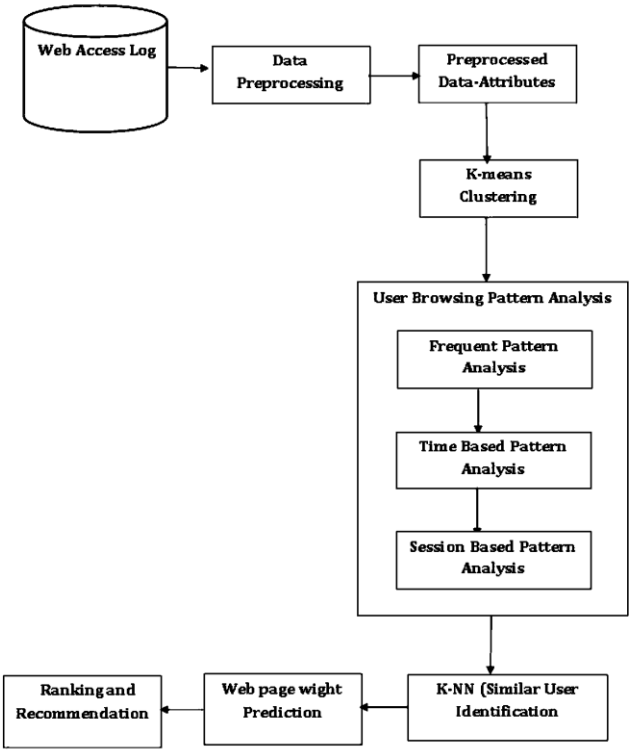


Figure 1. Architecture Dia

The main task of data pre-processing is to prune noisy and irrelevant data. The proposed web page recommendation system contains the K-means algorithm which is used to group of data according to the user IP address for finding the similar access patterns of the user sessions. Additionally for classification and prediction the KNN algorithm is implemented. The KNN algorithms to analyse data in distance based function and most nearest similar patterns are listed which is belongs from the other user therefore the proposed model also predicts the rarely accessed patterns. Thus to make the recommendations web usages data is personalized, based on URL frequencies, user navigational frequencies and time based data analysis. Additionally to combine these parameters a weighted technique is used. A combine weight for all the URLs is evaluated. According to the obtained weights the URLs are sorted and the maximum weight is selected as prediction of recommendation system. The highest weight shows the higher probability of visiting a web page after the current navigated web page. This project is contributed that improves the recommendation accuracy based on the session of user's web access. It provides more appropriate recommended web page to the active user.

3. Implementation

Modules:
Module 1: Data Preprocessing
Module 2: User Browsing Pattern Analysis
Module 3: Time based Browsing Pattern Analysis
Module 4: Similar User Browsing Pattern Analysis
Module 5: Next Web Page Recommendation
Modules Description:
Module 1: Data Preprocessing
Input: Raw Web Server Log File
Output: Preprocessed Log file

Web logs contains multiple records and information. When user enters on to the web, What data he accessed by him, how many times, Which websites are visited mostly, IP address of that users system and many more information. Web logs also contains the error or failure entries, some access records which are generated by search engine. Hence data preprocessing step perform data cleaning, formatting and grouping operation. In data cleaning all unwanted entries are removed and only that entries are extracted which are useful for recommendation operation.

Attributes Selection

During the pre-processing of log files the selected or targeted attributes are extracted and preserved in a database. These attributes are used for computing the different parameters on which the prediction of next web page is performed. It contains different kinds of attributes i.e. IP address, time stamp, requested URL, browser information and others. Among them some of the data is required for developing the proposed recommendation system and not all the attributes are used.

Module 2: User Browsing Pattern analysis
Input: Preprocessed Log file
Output: Frequent Pattern for each web log user
k-means Clustering

Each user accessed data is extracted using the K-mean clustering algorithm. k-means clustering is applied on the data to prepare group of data according to the user IP address. Each user IP address is represented as centroid in clustering. From the log file IP address from each entry and the corresponding access pattern is processed and merged to the closest centroid. Finally number of groups are obtained based on the IP address that contains the individual user' s web browsing pattern.

Frequent pattern analysis

The individual user's web browsing pattern is identified, find the most frequent accessed web pages for each user. The frequency of the individual web pages accessed by the user the following formula can be used.

Freq(Webpages)= 1 / N Σi=1toN (pagecounti)
N-Total number of pages accessed by specific user
Pagecounti -URL Frequency

Module 3: Time based Browsing Pattern Analysis
Input: Preprocessed Log file
Output: URL Access Time and URL Session for each web log user

The preprocessed log data, the time stamp is analyzed for a individual user browsing pattern. The amount of time spent on a particular web page is calculated using the following formula,

Time(Webpages)= 1 / N Σi=1toN (webpagetimei)
N-Total amount of time web pages accessed by specific user
webpagetimei-URL Accessing Time

Session based Browsing Pattern Analysis

In this module the session based navigational pattern is analyzed. A session is a list of web pages accesses from a given user during a period of time. For the task of identifying the list of web pages visited during a user's session at morning, afternoon or evening likewise. It provides more appropriate recommended web page to the active user.
Session Formula
Session1(Webpages)=1/N1Σi=1toN1 (sessionpagecounti)
N1-Total number of pages accessed by session1
sessionpagecounti-URL Frequency for session1
Session2(Webpages)=1/N2Σi=1toN2 (sessionpagecounti)
N2-Total number of pages accessed by session2
sessionpagecounti-URL Frequency for session2
Session3(Webpages)=1/N3Σi=1toN3 (sessionpagecounti)
N3-Total number of pages accessed by session3
sessionpagecounti-URL Frequency for session3

Module 4: Similar User Browsing Pattern Analysis

Input: Frequent Pattern of Web log User
Output: Extract Similar user Browsing Pattern

The k-NN classification algorithm to identify the target users search pattern, matching it to a web logs user group. It takes target user previous logs (frequent pattern) as a input and find out which user access the same pattern, from that data it predicts the users interest. The neighbors (similar user) of target user browsing pattern is evaluated by measure the Euclidean Distance between the target user frequent pattern and all the web log user frequent pattern. From k-minimum distance from web log user, most nearest pattern of active user will extracted.

Module 5: Next Web Page Recommendation
Input: Nearest Browsing Pattern
Output: Recommended the Next Web Page
Multi-Attribute URL Weight Prediction

From similar nearest frequent pattern are identified, the strength of next upcoming URL is computed based the multi-attribute browsing patten. The multi-attribute parameters namely user navigational frequency, time based URL and session based URL to combine weight for all the URLs are evaluated.

Weight(Webpages)= w1*Freq(Webpages)+w2*Time(Webpages) + w3* session(Webpages)

These weights are further sorted and by the rank of weights the next most possible web page is predicted. According to the current user input pattern system generate the prediction of next web page.

4. Algorithm

1. Start.

2. System will access the log files of user and do preprocessing to find most visited web pages and also removing an error entries from log files(error such as 404 page not found, connection failed, internal server error etc.).

3. . Finds the frequent pattern of web logs user using formula:

Frequency (Webpages) =1/ N Σi=1toN (pagecounti)

N-Total number of pages accessed by specific userPagecounti-URL Frequency

4. Calculation of URL Access Time and URL Session for each web log user

Time (Webpages)=1/NΣi=1toN(webpagetimei)

N-Total amount of time web pages accessed by specific userwebpagetimei-URL

Accessing TimeSession Formula

Session1(Webpages)=1/N1Σi=1toN1(sessionpagecounti)

N1-Total number of pages accessed by session1sessionpagecounti-URL Frequency for session1

Session2(Webpages)=1/N2Σi=1toN2(sessionpagecounti)

N2- Total number of pages accessed by session2 Sessionpagecounti-URL Frequency for session2.

Session3(Webpages)=1/N3Σi=1toN3(sessionpagecounti)

N3-Total number of pages accessed by session3 sessionpagecounti-URL Frequency for session3

5. Calculate the weight of webpages and using formula: Weight

Webpages)=w1*Freq(Webpages)+w2*Time(Webpages)+w3* session(Webpages)

6. Using weight (webpages) recommend the web pages to the users.

7. End.

III. RESULTS AND DISCUSSION

Training Time

The amount of time consumed during the training of the system is termed as the training time of the algorithm. Graph.1.1 shows the training time of the algorithms in terms of milliseconds.

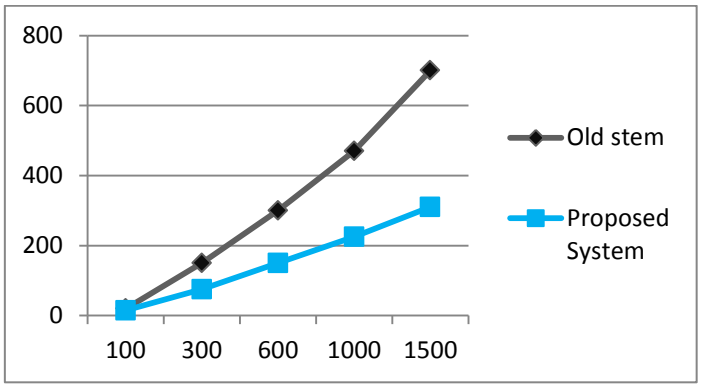
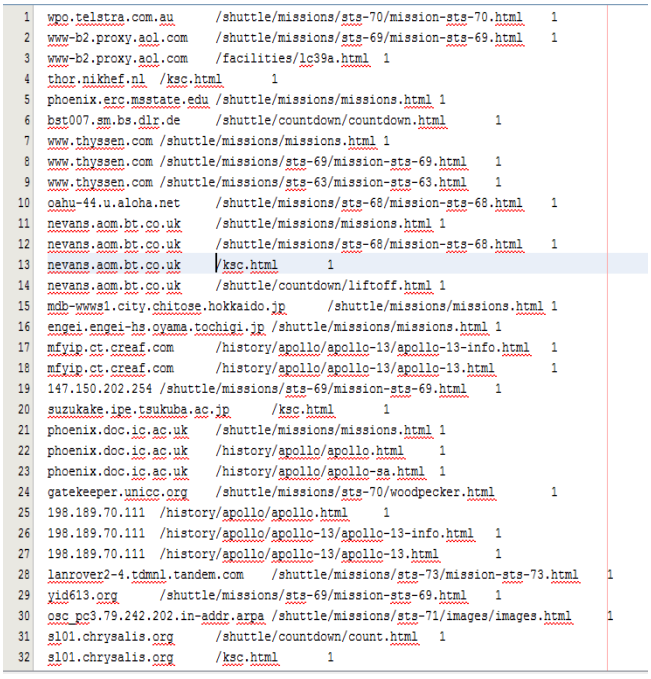


Figure 2. Training Time of datasets

Figure 3 and Figure 4. Show the classified data file of project.



```
wpo.telstra.com.au      /shuttle/missions/sts-70/mission-sts-70.html    1
www-b2.proxy.aol.com    /shuttle/missions/sts-69/mission-sts-69.html    1
www-b2.proxy.aol.com    /facilities/lc39a.html  1
thor.nikhef.nl  /ksc.html       1
phoenix.erc.msstate.edu /shuttle/missions/missions.html 1
bst007.sm.bs.dlr.de     /shuttle/countdown/countdown.html       1
www.thyssen.com /shuttle/missions/missions.html 1
www.thyssen.com /shuttle/missions/sts-69/mission-sts-69.html    1
www.thyssen.com /shuttle/missions/sts-63/mission-sts-63.html    1
oahu-44.u.aloha.net     /shuttle/missions/sts-68/mission-sts-68.html    1
nevans.aom.bt.co.uk     /shuttle/missions/missions.html 1
nevans.aom.bt.co.uk     /shuttle/missions/sts-68/mission-sts-68.html    1
nevans.aom.bt.co.uk     /ksc.html       1
nevans.aom.bt.co.uk     /shuttle/countdown/liftoff.html 1
mdb-wwws1.city.chitose.hokkaido.jp      /shuttle/missions/missions.html 1
engei.engei-hs.oyama.tochigi.jp /shuttle/missions/missions.html 1
mfyip.ct.creaf.com      /history/apollo/apollo-13/apollo-13-info.html   1
mfyip.ct.creaf.com      /history/apollo/apollo-13/apollo-13.html        1
147.150.202.254 /shuttle/missions/sts-69/mission-sts-69.html    1
suzukake.ipe.tsukuba.ac.jp      /ksc.html       1
phoenix.doc.ic.ac.uk    /shuttle/missions/missions.html 1
phoenix.doc.ic.ac.uk    /history/apollo/apollo.html     1
phoenix.doc.ic.ac.uk    /history/apollo/apollo-sa.html  1
gatekeeper.unicc.org    /shuttle/missions/sts-70/woodpecker.html        1
198.189.70.111  /history/apollo/apollo.html      1
198.189.70.111  /history/apollo/apollo-13/apollo-13-info.html    1
198.189.70.111  /history/apollo/apollo-13/apollo-13.html         1
lanrover2-4.tdmnl.tandem.com    /shuttle/missions/sts-73/mission-sts-73.html    1
yid613.org      /shuttle/missions/sts-69/mission-sts-69.html    1
osc_pc3.79.242.202.in-addr.arpa /shuttle/missions/sts-71/images/images.html    1
sl01.chrysalis.org      /shuttle/countdown/count.html   1
sl01.chrysalis.org      /ksc.html       1
```

Figure 3. Classified data file.

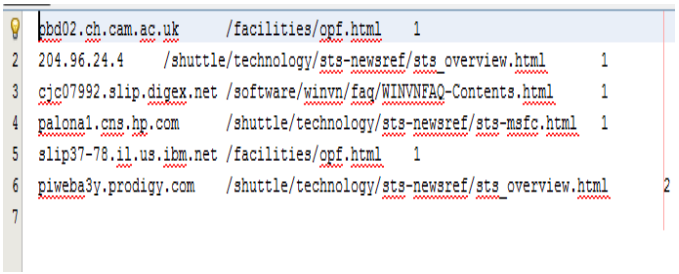


Figure 4. Similar Neighbour data file.

IV. CONCLUSION

Data preprocessing is an important task of Web log mining application. Therefore, data must be processed before applying data mining techniques to discover user access patterns from web log. The Proposed web recommendation system is concept by which the previous or historical user navigation data is analyzed and based on the most likely navigated technique the next web page access is predicted. This project is contributed that improves the recommendation accuracy based on the session of user's web access. It provides more appropriate recommended web page to the active user we will recommend to implement the existing system using association rule mining technique to more accurate result of next web page recommendations system.

V. REFERENCES

[1] Arvind Verma, Balwant Prajapat, "User Next Web Page Recommendation using Weight based Prediction" , International Journal of Computer Applications (0975 8887) Volume 142, No. 11, May 2016.

[2] K. Srinivas, P. V. S. Srinivas, A. Govardhan, V. Valli Kumari, "Periodic Web Personalization for Meta Search Engine", IJCST Vol. 2, Issue 4, Oct-Dec. 2011

[3] Neha Sharma & Pawan Makhija, Web usage Mining: A Novel Approach for Web user Session Construction, Global Journal of Computer Science and Technology: E Network, Web & Security, Vol. 15, Issue 3, 2015.

[4] Haidong Zhong, Shaozhong Zhang, Yanling Wang, Shifeng Weng and Yonggang Shu, "Mining Users Similarity from Moving Trajectories for Mobile Ecommerce Recommendation, International Journal of Hybrid Information Technology Vol.7, No.4, pp.309-320, 2014.

[5] Zahid Ansari, A. Vinaya Babu, Waseem Ahmed and Mohammad Fazle Azeem, "A Fuzzy Set Theoretic Approach to Discover User Sessions from Web Navigational Data", IEEE Recent Advances in Intelligent Computational Systems.

[6] I. Petrović, P. Perković and I. Štajduhar, "A Profile- and Community-Driven Book Recommender System, 38th International Convention on Information and Communication Technology, Electronics and Microelectronics(MIPRO), 2015.

[7] Lina Yao and Quan Z. Sheng, Aviv Segev, Jian Yu, "Recommending Web Services via Combining Collaborative Filtering with Content-based Features", IEEE 20th International Conference on Web Services,2013.

[8] Quanyin Zhu, Hong Zhou, Yunyang Yan, Jin Qian and Pei Zhou, "Commodities Price Dynamic Trend Analysis Based on Web Mining", Third International Conference on Multimedia Information Networking and Security, 2011.

VI. ABOUT THE AUTHORS

Prof. Umesh A. Patil
Assistant Professor & Head of the Department,
Computer Science and Engineering
D. Y. Patil Technical Campus,
Faculty of Engineering & Faculty of Management,
Talsande, Kolhapur,India.
Area of interest is Discrete mathamatics of computer,Theory of computation, Compiler. Data Structure e.t.c.

Mr.Avinash Kunnure.
UG Student,
Computer Science and Engineering
D. Y. Patil Technical Campus,
Faculty of Engineering & Faculty of Management,
Talsande, Kolhapur,India
Area of interest is Theory of computation, Compiler, Programming language ,Data Structure.e.t.c.

Mr.Vaibhav Herwade.
UG Student,
Computer Science and Engineering

D. Y. Patil Technical Campus,
Faculty of Engineering & Faculty of Management,
Talsande, Kolhapur,India
Area of interest is Theory of computation, Compiler,
Programming language ,Data Structure.e.t.c.

Mr.Vijaykumar Dawarpatil.
UG Student,
Computer Science and Engineering
D. Y. Patil Technical Campus,
Faculty of Engineering & Faculty of Management,
Talsande, Kolhapur, India
Area of interest is Theory of computation, Compiler,
Programming language ,Data Structure.e.t.c.

Mr. Satyawan Magar
UG Student,
Computer Science and Engineering
D. Y. Patil Technical Campus,
Faculty of Engineering & Faculty of Management,
Talsande, Kolhapur,India
Area of interest is Theory of computation, Compiler,
Programming language ,Data Structure.e.t.c.