International Journal of Multidisciplinary Research and Development

International Journal of Multidisciplinary Research and Development

Volume: 2, Issue: 6, 426-430
June 2015
www.allsubjectjournal.com
e-ISSN: 2349-4182
p-ISSN: 2349-5979
Impact Factor: 3.762

Jothish Chembath
Ph.D Research Scholar, Department of Computer Science, Karpagam University, Coimbatore.

S.K. Mahendran
Director, SVS Institute of Computer Applications, Coimbatore, India.

Correspondence:
Jothish Chembath
Ph.D Research Scholar, Department of Computer Science, Karpagam University, Coimbatore.

Transaction Identification Algorithm Enhanced With User Pruning and Combined Maximal Forward Reference and Reference Length Approach for Improving Prediction of Next Web Page from Web Log Entries

Jothish Chembath, S.K. Mahendran

Abstract
Next web page prediction is one of the vital operations performed by web masters to improve customer loyalty and reduce retention. It consists of various steps like preprocessing, pattern discovery and analysis and the focal point of this work is preprocessing. The steps in pre-processing include cleaning, user identification, session identification and transaction identification. In this work, a transaction identification algorithm is proposed, which segments a session into sequence of meaningful pages. The proposed algorithm combines techniques like pruning of irrelevant users to reduce web log data size, combined Maximal Forward Reference and Reference length approach for estimating automatic cutoff time for session and transaction identification. An algorithm that performs path completion on the transactions identified is also presented. Experimental results prove that the combined application of these algorithms increase both preprocessing task and prediction of next web page.

Keywords: Next Page Prediction, Path Completion, Preprocessing, Transaction Identification, Pruning, Web Log Data, Web Usage Mining

1. Introduction
The tremendous growth of World Wide Web (WWW) have motivated several researches to focus on studies that analyze details regarding the user's interactions and behaviors in order to identify their browsing patterns and preferences. This knowledge is then used to improve their experience during a transaction and to better serve them (Tahil and McArdle, 2011) [13]. User details, stored in 'Weblogs', are generally used for this purpose. The result of such analysis presents knowledge of users' intentions, are used in several manners by web masters. These techniques have the goal of saving browsing and searching time, while at the same time reducing the retrieval time and bandwidth load on network and are termed as Next Web Page Prediction (NWPP) System. Several researchers have focused on developing systems for this purpose (Badhe and Shirsat, 2013; Suguna and Sharmila, 2013) [2, 11]. But owing to the dynamic nature of the WWW and e-commerce industry, the research in this area is still very active and focuses on identifying techniques that improve these systems in terms of accuracy and speed.
The general steps involved in NWPP system are preprocessing, pattern mining and pattern analysis. This paper focuses on the first step (Preprocessing), which when handled correctly can improve the performance of NWPP system. The purpose of preprocessing is to transit various input such as content, structure and usage information into the format which data mining algorithms can handle easily (Han *et al.*, 2001) [6]. This paper is oriented towards the preprocessing stage of NWPP system and presents a Transaction Identification Algorithm enhanced with Pruning and Combined Maximal Forward Reference Approach (MFRA) and Reference Length Approach (RLA). The rest of the paper is organized as follows. Section 2 presents the steps in preprocessing step of NWPP with emphasis to the task of transaction identification. Section 3 discusses the results obtained during performance evaluation, while Section 4 concludes the work with future research directions.

2. Methodology
In preprocessing, the main focus here is to retain only useful data from the raw web log and to format in a way that can be easily used by the prediction model. The preprocessing stage performs various tasks for this purpose as listed below.

- Cleaning of web log data - Removes unwanted and irrelevant entries in web log data
- User Identification - User identification is the process used to identify all the unique users accessing a web site or portal.
- Session Identification - A user session is defined as a sequence of requests made by a single user over a certain navigation period
- Transaction Identification - Used to identify the complete access path of each user for each session

This paper is oriented towards enhancing the transaction identification step and the proposed algorithm is referred to as Transaction Identification Algorithm enhanced with Pruning and Combined MFRA and RLA (TIAPC).

2.1. Cleaning Algorithm

The data removed during cleaning are not important for user navigation prediction and hence can be deleted safely from the log file. Initially, the cleaning algorithm removes all unwanted (example - images, java scripts, flash animations, video, etc.) and redundant data (entries made by repeated access to the same page by the same user) from web log data files. The cleaning algorithm also removes accesses made by non-humans (examples - entries made by web crawlers and Spiders). The algorithm also removes erroneous references (failed page requests), which can be identified using the status attributes in the web log data.

2.2. User Identification Method

The second step of preprocessing is User Identification (UI), which is the task of identifying unique users of a website. Currently this performed using various methods. Examples include cookies (Eirinaki and Vazirgiannis, 2003; Huysmans *et al.*, 2003) [5, 7], Identd protocol (RFC 1413, 2010), unique user names and IP address. The usage of IP address for unique user identification is the most frequently used method as it is simple, easy to capture and is never empty. Hence this method is used in this work

2.3. Session Identification Algorithm

A user session is defined as a sequence of requests made by a single user over a certain navigation period. The third step, Session Identification, is used to segregate the page accesses of each user into individual sessions. In general, there are two methods that are frequently used for session identification. They are either time-oriented or structure-oriented. This research work uses a time-oriented approach, whose steps are shown in Figure 1. In this algorithm, in general, TT is assigned a value 30 minutes. In this research work, it is automatically calculated using the procedure described in the next section (Transaction Identification Algorithm).

- Let θ_1 be the time stamp of the first request, R_1.
- A new session (S) is started at this time (θ_1)
- Repeat
  - Let θ_2 be the time stamp of the next request, R_i
  - Add R_i to S
  - Till $\theta_1 - \theta_2 <$ Time Threshold (TT).

Fig 1: Session Identification Algorithm

2.4. Transaction Identification Algorithm

The main goal of Transaction Identification Algorithm (TIA) is to identify, from the session information and user information, individual needs of the user in the same session. A transaction consists of more than one page (called as a sequence of pages), but not all pages of a session. The TIA divides and identifies such meaningful transactions and takes into account the amount of time spend by the users on each page of session. Depending upon this time, a web page can either be a content page or an auxiliary page. Whenever a content page is encountered, the set of pages are sequenced as a single transaction and the algorithm begins to build a new transaction. However, this simple method has issues as given below.

(i) The presence of irrelevant users who have no actual interest on the website do not contribute much to TIA and increases time complexity.
(ii) The content and auxiliary pages are found using a cutoff time estimated as the difference between the time of next reference and the current reference (Cooley, 1997) [4]. This estimation is not suitable for all web sites, as it ignores (i) time related to the size of data being transferred and (ii) data transfer rate.
(iii) The identified sequence in a transaction may have gaps (incomplete sequence), which degrades the prediction performance.

This paper proposes a TIA algorithm that solves all the above issues. The proposed algorithm is termed as TIAPC consists of the steps listed below.

A. Perform Grouping of Users to identify relevant and irrelevant users
B. For each session identify content pages and estimate cutoff time
C. Session Identification using the above cutoff time
D. Construct transactions
E. Perform path completion for incomplete transaction

A. Grouping of Users

As details regarding irrelevant users are not required during prediction, the TIAPC algorithm removes these entries from the session. For this purpose, an Integrated Clustering and Classification System for Grouping Users (I2CGU) is proposed (Figure 2). The clustering algorithm used is K-Means algorithm and grouping is performed using C4.5 decision tree classifier. While creating decision rules for C4.5 decision tree classifier, five attributes, namely, total session time (SeT), total time the user stays at the site (T), total number of accessed pages during the whole session (N), access methods (AM) used to interact with the site and depth wise access from a particular page are used to recognize relevant and irrelevant users. Each of the attributes were assigned parameter values, 15-30 minutes, > 30 seconds, > 5 pages, GET and POST method, Depth wise Reference (DR) by the user in a particular section respectively. These values are the same as used by Suneetha and Krishnamoorthi (2010) [12]. This system is referred to as RGU algorithm and uses decision rule-based algorithm using C4.5 for user grouping. The decision rules are formulated as given in Table 1.

B) ACTE Algorithm

Information regarding the type of web page is exploited to calculate the cutoff time during Session Identification. For this purpose, there exist two methods, namely, Reference Length Approach (RLA) or Maximal Forward Reference Approach (MFRA). The RLA produces better results when compared with MRFA if the content pages are identified correctly (Cooley *et al.,* 1999) [3].

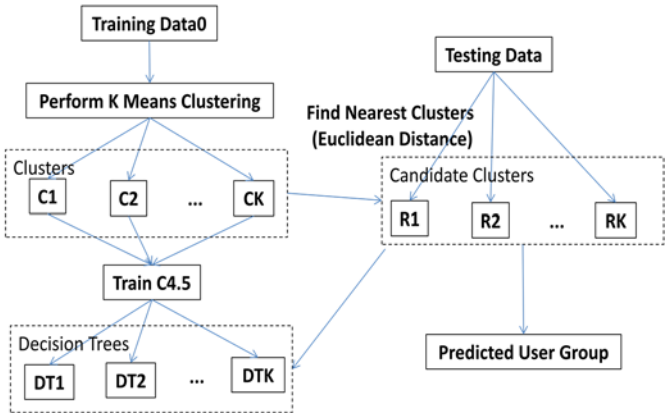

Fig 2: Steps in I2CGU

Table 1: Decision Rules

Rule No.	Description	User Grouping
R1	T < 30 and N < 5 and Method used GET and DR = 'NO'	Irrelevant
R2	T < 30 and N < 5 and Method used GET and DR ='YES'	Relevant
R3	T < 30 and N < 5 and Method used POST	Relevant
R4	T < 30 and N > 5 and Method used GET and DR ='NO'	Irrelevant
R5	T < 30 and N > 5 and Method used GET and DR ='YES'	Relevant
R6	T < 30 and N > 5 and Method used POST	Relevant
R7	T > 30 and N < 5 and Method used GET and DR = 'NO'	Irrelevant
R8	T > 30 and N < 5 and Method used GET and DR ='YES'	Relevant
R9	T > 30 and N < 5 and Method used POST	Relevant
R10	T > 30 and N > 5 and Method used GET and DR ='NO'	Irrelevant
R11	T > 30 and N > 5 and Method used GET and DR ='YES'	Relevant
R12	T > 30 and N > 5 and Method used POST	Relevant

For this reason, in this work, the MRFA algorithm is first used to identify the content pages, which is then used to estimate the cutoff time, used as threshold in RL to identify the sessions and transactions. The steps involved are shown in Figure 3, where P_c is the percentage of content pages in S. Next, the cutoff time between auxiliary and content pages is obtained using the Equation in Step 5 of Figure 3.

Step 1 : for i = 1 to n
  if ($ReferURL_i$==$ReferURL_{i-1}$) or
    $ReferUL_{i-1} \in C$
  Add i to C
 end for
Step 2 : Set C_N = size of C
Step 3 : $P_C = (N/C_N)*100$
Step 4 : Set MRF = Mean RL of all pages in W
Step 5 : C_oT = -MRF * ln P_c

Fig 3: ACTE Algorithm

C) Session Identification

The cutoff threshold estimated in ACTE algorithm (C_oT) is then assigned to TT in Figure 1 to identify the sessions.

D) Construct Transactions

The C_oT threshold estimated in the previous step is then with RLA to identify the auxiliary pages and the optimal page sequences for the transactions in the session. Figure 4 presents the steps of this algorithm. The conventional RLA finds the reference length as the difference between the access time of the next and the present page. However, during browsing, this difference alone will not be efficient, as the time considered by RLA includes time taken to transfer the data, launching of applications (example : online audio or video files). Thus, estimation of time should be taken into account not only the access time but also the transfer rate. Thus, the RL is modified (Equation 1) and is used to find users' access patterns.

$$RefLen = (mod(Date(URL(i))- Date(URL(i-1))) - bytes\_sent / c \quad (1)$$

Let A be the set of auxiliary pages whose RefLen > C_oT
LastSessionID = 0
CRF = NULL; // Current Transaction's ReferURL
Let FTS = NULL; // FinalTransactionSet
 For i = 1 to n // i ∈ A
  if SessionID(i) == LastSessionID
   RefLen = -1
   Tr = <SessionID(i), Date, RefLen>
   Insert Tr into FTS
  end
  CRU = $ReferURL_i$; //Current ReferURL
  If CRU == $ReferURL_{i-1}$ then Insert i to FTS
  if URL(i) == URL(i-1)
   delete i from S
  else
   RefLen = (mod(Date(URL(i))- Date(URL(i-1))) - bytes_sent / c
   Tr = <SessionID(i), Date, RefLen>
   Insert Tr into T
  end
end for

Fig 4: MRL Approach

In the above algorithm, W refers to cleaned web log data and S be the set of users sessions, that is, S = {SessionID, <URL_1, $ReferURL_1$, $Date_1$>, …, <URL_n, $ReferURL_n$, $Date_n$>} where 'n' denotes the number of transactions in S and $1 \le n \le N$ (N denotes number of transaction in the pruned web log data). A Transaction T consists of a set of pages accessed by the users, that is, T = {SessionID, <URL_1, $Date_1$, $RefLen_1$>, …, <URL_n, $Date_n$, $RefLen_n$>} where RefLen (reference length) refers to the time spent on page URL. Let C be the set of content pages of the form <URL_1, …, URL_n}. Let A be the set of auxiliary pages of the form <URL_1, …, URL_n} whose RefLen > cutofftime. CRF refers to the Current Transaction's ReferURL while FTS refers the Final Transaction Set.

E) Path Completion in Incomplete Transactions

The transactions obtained in this stage some transactions may contain missing URLs, which results in incomplete sequences. To attain complete transactions, the path completion algorithms are used. This facilitates complete user access details, which in turn, will improve prediction performance. Path completion is considered as a critical step of preprocessing, which point towards the situation where the number of URLs (Uniform Resource Locators) is less than the actual URLs browsed by the user. Incomplete paths in weblog

data files can occur due to (i) the caching problems in proxy servers, there are possibilities of pages missing after the construction of transactions in weblog files (Li and Feng, 2009) and/or (ii) Failure to record requested pages can also occur during local buffering (Anand and Aggarwal, 2012). The path completion algorithm used in this work consists of the steps given below.

- After identifying path for each session, if any of the URL specified in the Referrer URL is not equal to the URL in the previous record then that URL in the Referrer URL field of current record is inserted into this session and thus path completion is obtained.
- The next step is to determine the reference length of new appended pages during path completion and modify the reference length of adjacent ones.
- The reference length of adjacent pages is also adjusted.

3. Experimental Results

In order to evaluate the performance of the preprocessing algorithms, web log dataset from NASA Kennedy Center Space (http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html) is used. This dataset has log entries collected in the period of 01-07-1995 to 31-08-1995, occupying 205.2MB storage space in uncompressed form. It has a total of 3,461,612 log entries. The performance measuring factors used for evaluating the preprocessing algorithm is the percentage of reduction obtained on number of transaction and memory are used. The experiments also analyze the effect of these algorithms on prediction of next web page. Three performance metrics, namely, accuracy, coverage and F1 measure are used for this purpose. The prediction model used in this stage was proposed by Jalali *et al*. (2010) [8] and is referred to as LPA (Longest common sequence-based Prediction Algorithm).

The effect of using cleaning algorithm on raw web log data with respect to number of transactions and storage size are given in Figures 5 and 6 respectively.

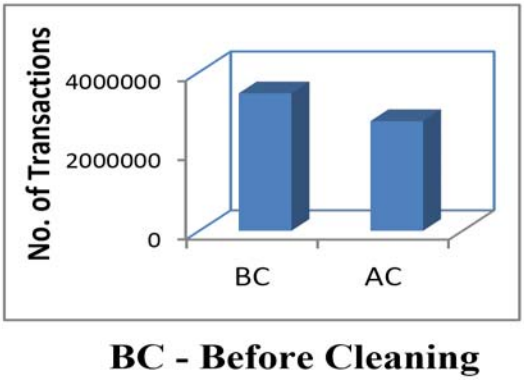

Fig 5: Effect of Cleaning (No. of Transactions)

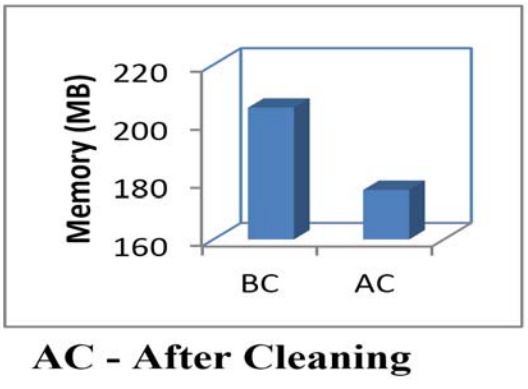

Fig 6: Effect of Cleaning (Storage Size)

From figures 6 and 7, it is evident that the application of preprocessing on raw web log data has a positive impact on both web log file size and storage usage. An efficiency gain of 20.3% with respect to number of transactions and 13.8% with respect to storage usage were obtained while using cleaning algorithm in NWPP system.

The effect of I2CGU on number of transactions and storage space is shown in Figures 7 and 8 respectively. The efficiency of the user grouping algorithm is evaluated using the accuracy parameter and the results obtained are shown in Figure 9.

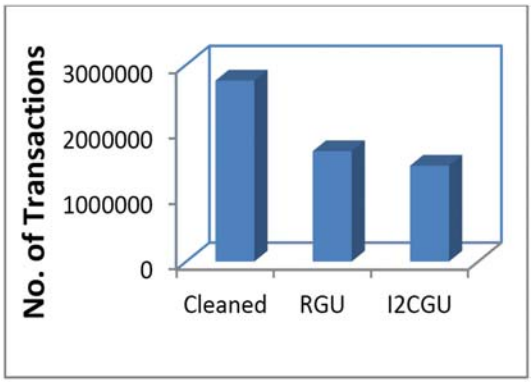

Fig 7: Effect of User Grouping (No. of Transactions)

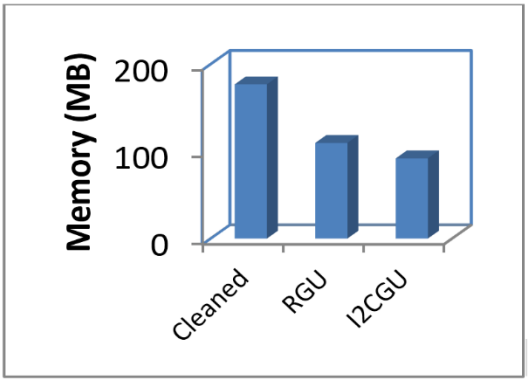

Fig 8: Effect of User Grouping (Storage Size)

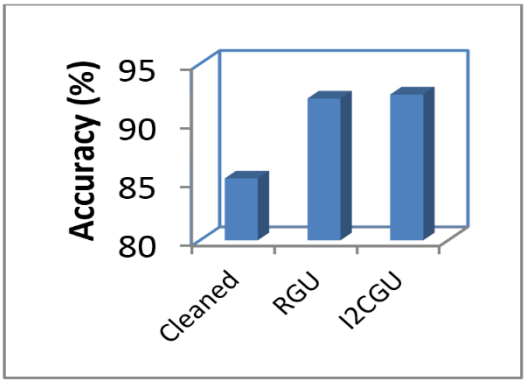

Fig 9: Accuracy (%) of User Grouping Algorithm

The I2CGU algorithm was able to reduce the web log data by 47% and 48% reduction with respect to number of transactions and storage space utilized respectively, while the RGU algorithm reduced to only 39% and 38% respectively. From Figure 9, it is clear that the proposed I2CGU can identify relevant and irrelevant users accurately when compared to RGU algorithm. This proves that the inclusion of clustering algorithm to train C4.5 decision tree classifier has a positive impact on identifying the type of user browsing the website. Thus, the results prove that the inclusion of K-Means

algorithm to group homogeneous pages together before applying c4.5 classifier is effective in finding the relevant and irrelevant users in web log data in terms of accuracy, number of transaction and storage space occupied.

The coding scheme presented in Table 2 is used during the discussion that analyzes the effect of preprocessing on prediction of next web page. Table 3 presents the accuracy, coverage and F1-Measure of the prediction model with and without applying the preprocessing algorithms.

Table 2: Coding Scheme Used

Description	Code	Description	Code
LCS-Based Prediction Algorithm	LPA	LPA with TIA Based on MFRA	LPA-MFRA
LPA with TIA Based on RLA	LPA-RLA	TIAPC with Path Filling	TIAPC

Table 3: Effect of Preprocessing on Prediction

Prediction Model	Accuracy	Coverage	F1 Measure
LPA	85.62	1.6325	0.8241
LPA-RLA	87.58	1.3743	0.6926
LPA-MFRA	86.72	1.3959	0.7036
TIAPC	90.35	1.2184	0.6133

From Table 3 results, it is clear that the proposed TIAPC is successful and has improved the performance of Longest Common Sequence-based Prediction Algorithm in terms of all three selected performance metrics. The prediction model incorporated with TIAPC model showed an accuracy gain of 5.24%, coverage gain of 25.57% and overall gain in terms of F1 Measure 25.36%, which further proves the importance of TIAPC in next web page prediction.

4. Conclusion

This paper presented an algorithm that improves transaction identification using pruning irrelevant users, automatic identification of cutoff time for both session and transaction identification and path completion algorithm. The cutoff time is estimated by using MFRA first, which is then used by RLA to identify transactions. The transactions thus identified face the issue of being incomplete, which is solved by using a path completion algorithm. Experimental results prove that the proposed enhancements for transaction identification have reduced number of transactions and storage space. Analysis on prediction has shown increased performance in terms of accuracy, coverage and F1 Measure, indicating that the reduction removes only irrelevant data from raw web log data and produces an optimal set of transactions that can be readily used by the prediction models. Future work includes analysis of various prediction models that can use the preprocessed data produced by TIAPC.

5. References

1. Anand, S. and Aggarwal, R.R. (2012) An Efficient Algorithm for Data Cleaning of Log File using File Extensions, International Journal of Computer Applications, Vol. 48, No, 8, Pp.13-18.
2. Badhe, N. and Shirsat, K.P. (2013) A Novel Approach for Web Recommendation System based on Sequential Access Patterns, IJAIS Proceedings on International Conference and workshop on Advanced Computing, Vol. 4, Pp. 36-40
3. Cooley, R., Mobasher, B. and Srivastava. J. (1999) Data preparation for mining world wide web browsing patterns, Knowl. Inf. Syst., Vol. 1, No. 1, Pp. 5–32.
4. Cooley, R., Mobasher, B., and Srivastava, J. (1997) Web mining : Information and Pattern Discovery on the World Wide Web, International Conference on Tools with Artificial Intelligence, Pp. 558-567.
5. Eirinaki, M. and Vazirgiannis, M. (2003) Web mining for web personalization, ACM Transactions on Internet Technology (TOIT), Vol. 3, Issue 1, Pp. 1-27.
6. Han, J. and Kamber, M. (2001) Data mining: concepts and techniques, Academic Press, San Diego, CA.
7. Huysmans, J., Baesens, B. and Vanthienen, J. (2003) Web Usage Mining: A Practical Study, Katholieke Universiteit Leuven, Dept. of Applied Economic Sciences.
8. Jalali, M., Mustapha, N., Sulaiman, N. and Mamat, A. (2010) WebPUM: A web-based recommendation system to predict user future movements, Expert Systems with Applications, Vol. 37, Pp. 6201-6212.
9. Li, Y. and Feng, B., 2009. The Construction of Transactions for Web Usage Mining. In the Proceedings of International Conference on Computational Intelligence and Natural Computing CINC'09, 1,121 -124
10. RFC 1413 (2010) Identification Protocol, http://www.rfceditor.org/rfc/ r
11. Suguna, P. and Sharmila, D. (2013) User interest level based preprocessing algorithms using web usage mining, International Journal on Computer Science and Engineering, Vol. 5 No. 09, Pp. 815-822.
12. Suneetha, K.R. and Krishnamoorthi, R. (2010) Classification of web log data to identify interested users using decision trees, International Conference on Computing, Communications and Information Technology Applications, (CCITA 2010), Coimbatore, India.
13. Tahir, G. and McArdle, M.B. (2011) Visualising user interaction history to identify web map usage patterns, 14th AGILE International Conference on Geographic Information Science, Advancing Geoinformation Science for a Changing World, Netherlands, Pp.1-7.