UNDERSTANDING USER'S NAVIGATION BEHAVIOR USING DATA MINING TECHNIQUES

Sampada K. Bari, Vaishnavi L. Bhole, Apeksha Patil, Surajkumar S. Varade
SSBT's College of Engineering and Technology,
Jalgaon, Maharashtra

Abstract : Data mining is the process of extracting knowledge from data. That knowledge is usually related with relationships between variables and/or observations and patterns likes clusters of observations. Discovering hidden patterns from the web logs is an upcoming research area. Predictions need to keep track of history data to analyze the usage behavior of the users. Web usage mining consists of pre-processing, pattern discovery and pattern analysis. 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. System emphasis is on the prediction of user behavior using web log file. This approach requires user session identification, clustering the sessions into similar clusters and developing a model for prediction using the current and earlier accesses. The system uses pre-processing of web log, hierarchical clustering technique and data mining algorithm for prediction.

IndexTerms - **User and Session identification, Levenshtein Distance, Hierarchical Clustering, Markov Model, Prediction**

I. Introduction

There are increasing research interests in using data mining techniques in prediction of user behavior.[1] The amount of information in the web is increasing every day and the demand for the information is also proportionally increasing. Websites provide web log files which are automatically created and maintained by the web servers. Every access to the website, including each view of the HTML document, image or other object is logged.

1.1 Web Mining

Web mining is application of data mining techniques to discover patterns from the web. As the name suggests, information is gathered by mining the web means extracting the useful and valuable information or knowledge from the huge distributed and unstructured data. [2] It has 3 types:

1. Web Content Mining: It is used to examine data collected by web spiders and search engines.
2. Web Structure Mining: It is used to examine data related to the particular website's structure.
3. Web Usage Mining: It mines the usage Patterns on the web and improves web usability and user experience.

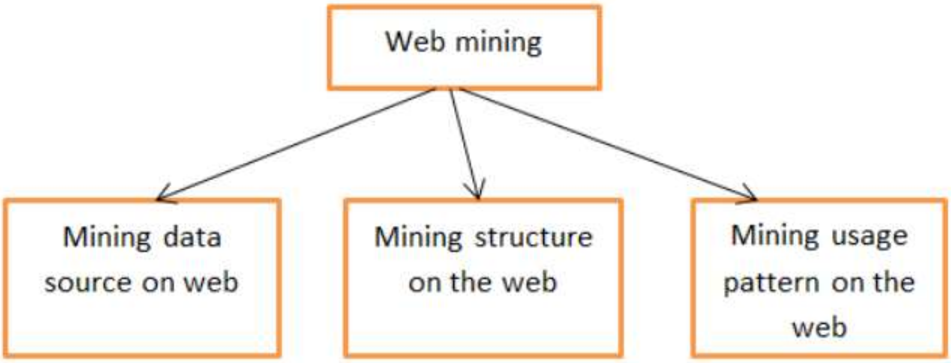

Web mining
Mining data source on web
Mining structure on the web
Mining usage pattern on the web

Figure 1.1: Types of Web Mining

1.2 Web Log

The raw web log file format is one line of text for each access to the website. A Web log file records activity information when a Web user submits a request to a Web Server. This contains information about who visited the site, where they came from, what they were doing on the website and how long they have accessed the page. Web log file are of two types: Common log format and Extended log format or Combined log format.

1.2.1 Common Web Log Format

The common web log format has the fields such as Client IP Address, User id, Access Date and Time, HTTP Request Method, Path of the Resource on the web server, Protocol used for Transmission, Status code for the Transmission and Number of bytes transmitted.

1.2.2 Extended Web Log Format

The extended web log format has some additional fields like referrer and user agent. Referrer: It is the link from where the user has reached at the required web site. User Agent: It is the browser through which the user accesses the web site.

E.g.: 202.244.227.66 - - [01/Jul/1995:05:25:36 -0400]"GET http://www.enggresources.com/shuttle/missions/missions.html HTTP/1.0" 200 12054 http://www.google.com Mozilla.

In above example,

- 202.244.227.66 is a Client IP Address.
- 01/Jul/1995 is an Access Date.
- 05:25:36 is an Access Time.
- HTTP is a protocol user for Transmission.
- http://www.enggresources.com/shuttle/missions/missions.html is the Path of the Resource on the web server.
- 0400 is a user ID.
- 200 is a status code for Transmission.
- 12054 is a Number of bytes transmitted.
- http://www.google.com is a referrer.
- Mozilla is a user agent.

1.3 User and Session Identification

Identifying a user by observing IP-addresses is referred as user identification. By comparing IP-addresses, we can identify users. If there are same IP-addresses, then the user is same; otherwise the user is new.

Now, a series of page views of a single user when accessing the whole web is called as a session. The purpose of session identification is to separate the web logs of one user into individual sessions for further analysis. The method for identification of session is to determine some threshold. [3] This threshold is nothing but the time period. If this threshold is exceeded by the time period between accessing pages, then it is considered as the user starts a new session. In this system, this threshold is determined as 30 minute.

1.4 Motivation

At the time of need, it is not easy to get information relevant to the topic from this huge amount of data. To increase speed and performance of browsing, recommendation is needed. Recommendation is nothing but prediction of anticipating pages that user may visit so that user can get more information about his/her domain of search easily and speed of their surfing is improved. Page prediction will predict where the user might visit next and preload that page ahead of time. [4] This makes the navigation faster. Page prediction involves analyzing and understanding the usage pattern for producing the useful information.

1.5 Problem Definition

Web Page prediction system is based only on the web access. To reduce the web page access latency, predicting the current user next move is essential. An integrated approach to predict the next page using pre-processing of web log, hierarchical clustering and data mining algorithm.

1.6 Hierarchical Clustering

Hierarchical clustering is one of the types of clustering. In hierarchical clustering, the data is not partitioned into a particular cluster in a single step. Instead, a series of partitions take place, which may run from a single cluster containing all objects to n clusters that each contain a single object or from n clusters containing a single data object to a single cluster containing n objects. Accordingly, it is subdivided into 2 methods:

1. Agglomerative method: it is processed by a series of fusions of the n objects into groups.
2. Divisive method: It separates n objects successively into finer groupings.

The proposed system uses the Agglomerative Hierarchical Clustering method. For clustering the sessions, measuring the similarity between the sessions is necessary. This is achieved by using an appropriate distance metric. There are different agglomerative techniques due to different ways of defining distance (or similarity) between clusters. Modified Levenshtein distance is used as the distance metric which is an improved form of Edit distance (Levenshtein Distance). Levenshtein distance does not consider page sequences but modified Levenshtein distance technique takes page sequence into account. It's the best technique for Hierarchical clustering of web sessions.

1.7 Markov Model

A Markov Model is a model used to model systems that changes randomly. It is considered that future states depend only on the current state, not on the previously happened events. This consideration is known as Markov property. First order Markov model gives the probability of the next state using the current state only without considering the previous states. It is very difficult to predict using only current state. A little bit of past memory is necessary to predict the future state. This proposed system uses

Higher Markov Model. It takes current state as well as previous states into account for future prediction. It gives probabilistic values about future states.

II. RELATED WORK

Virendra R. Rathod et al., in [4] has used two different clustering techniques as Fuzzy C-means clustering algorithm and Markov Model has investigated to predict the web page.

The server logs from the MSNBC dataset has been used by R Geetharamani et al., in [5] for their research. They has used Apriori Prefix Tree(PT) algorithm for prediction of probable subsequent page in the usage of web pages listed in the MSNBC dataset based on their navigating behavior.

P. G. Om Prakash et al., in [6] has classified the data of success response and analyze the user navigation. The process of identification of user behavior consists of data collection, query parser, pre-processing and pattern analysis that will help to analyze and predict the user behavior in short time.

Anurag kumar et al., in [7] has studied the user behavior using web server log file prediction using web server log record, click streams record and user information. A Web log along with the individuality of the user captures their browsing behavior on a website and discussing the behavior from analysis of different algorithms and different methods like apriori algorithm and FP-growth algorithm.

K. R. Suneetha et al., in [8] has identified user behavior by analyzing web server access log file. They have used the in-depth analysis of Web Log Data of NASA website to find information about a web site, top errors, potential visitors of the site etc. which help to improve the system by determining occurred systems errors, corrupted and broken links by using web usage mining.

Sana M. Deshmukh et al., in [9] has proposed two algorithms. Modified Clustering Algorithm-I adapts to cluster users based on their similarity and Algorithm-II based on Preferred Path Mining Algorithm that performs path mining on the clusters found using algorithm-I. Outcome of Algorithm-II is in the form of prediction of web pages in future each interested user may visit.

III. Architecture and Modeling

Architecture of the system is illustrated in Figure 3.1. There are three models: pre-processing, hierarchical clustering and Markov model.

The system begins with extended log file as an input. Data cleaning algorithm is applied on the input file to clean data.

Algorithm: Data Cleaning.
Input: File containing log records.
Output: Cleaned and relevant records stored in log file.

```
1. For each record in an input file, read fields.3
        1.1. If fields have extensions like gif, jpg, css or have error code 404, 500, then remove records.
        1.2. Else store records in log file. – Until no record is remaining.
```

User and session identification algorithm is applied on this cleaned log file and set of sessions is obtained as an output.

Algorithm: Identification of user and session.
Input: File containing relevant log records.
Output: Set of sessions.

```
1. For each record in log file, repeat following steps until no entry is remaining in the log file.
2. Compare IP-address of one entry with IP-address of every other entry.
3. If any two IP-addresses match, then those both entries are considered to be from same user.
4. For each identified user, order all records by time and identify session by comparing each entry period less than equal to 30 minutes from web page's first entry and consider minimum only 5 pages in a session.
5. Repeat above step until no entry is remaining.
```

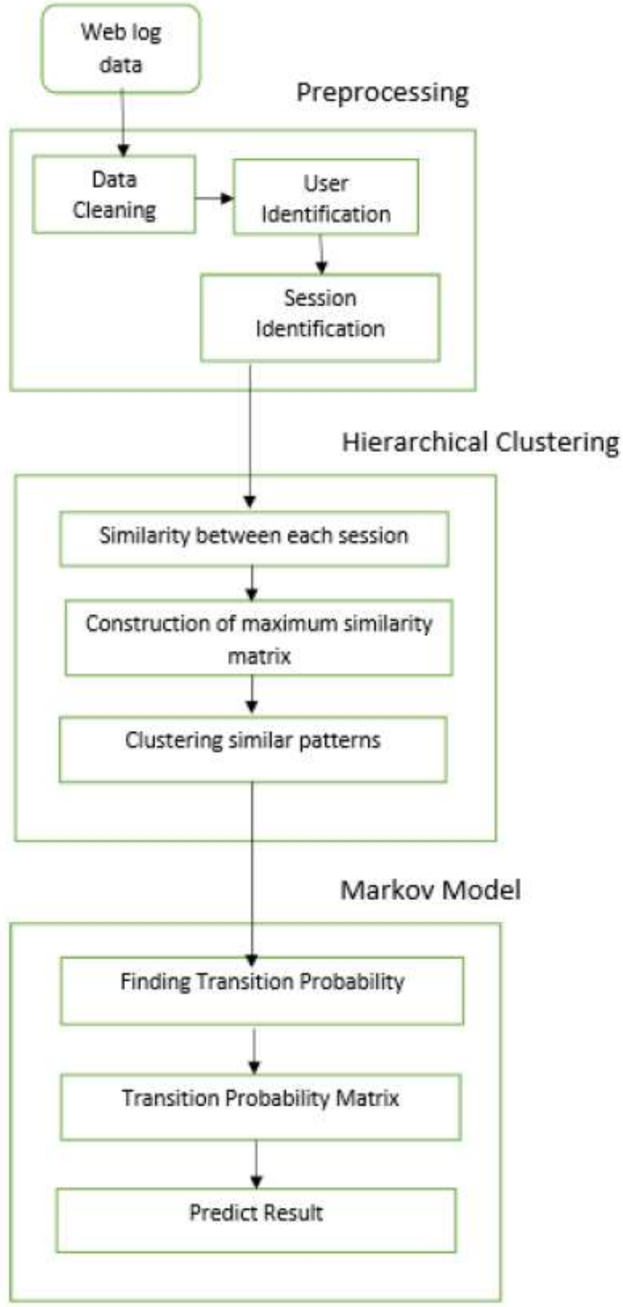

Web log data
Preprocessing
Data Cleaning
User Identification
Session Identification
Hierarchical Clustering
Similarity between each session
Construction of maximum similarity matrix
Clustering similar patterns
Markov Model
Finding Transition Probability
Transition Probability Matrix
Predict Result

Figure 3.1: Architecture

Set of sessions obtained are given to the Hierarchical Clustering algorithm and clusters of similar sessions are obtained.

Algorithm: Hierarchical Clustering.
Input: Set of sessions.
Output: k-clusters of similar sessions.

```
1. Construct Modified Levenshtein distance matrix between every pair of session using following equation:
   For 1<=i<=L1 and 1<=j<=L2,
        1.1. If $S_i = S_j$
         {
              $M(i,j) = Max \{ M(i-1,j-1)+2, M(i-1,j)-1, M(i,j-1)-1 \}$
         }
        1.2. Else
         {
              $M(i,j) = Max \{ M(i-1,j-1)-1, M(i-1,j)-1, M(i,j-1)-1\}$          ...(1)
         }
```

Where M (i, j) is maximum similarity score matrix. L1 and L2 are lengths of session S_i and Sj respectively.

2. Distance maintains order of navigation which is number of similar pattern.
3. Find the maximum number in that matrix.
4. Calculate the max similarity score using following formula:

$$\text{Max Similarity} = \text{max}/\ (L1+L2) \qquad \text{... (2)}$$

5. Merge similar patterns together and their sessions until no more merging possible.
6. Update the session list.

Higher Markov algorithm takes these clusters and gives transition probability of moving from the current page to the next page.

Algorithm: Higher markov model.
Input: Representative of cluster.
Output: Transition probability.
1. Compute transition probability using formula as follows:

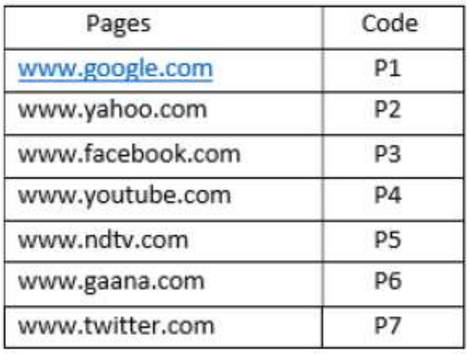
P [i,j]=Number of times there is transition from page$_i$ to page$_j$ / Number of times there is transition from page$_i$ to any other page. ... (3)

2. This probability is used for prediction.

IV. RESULTS AND DISCUSSION

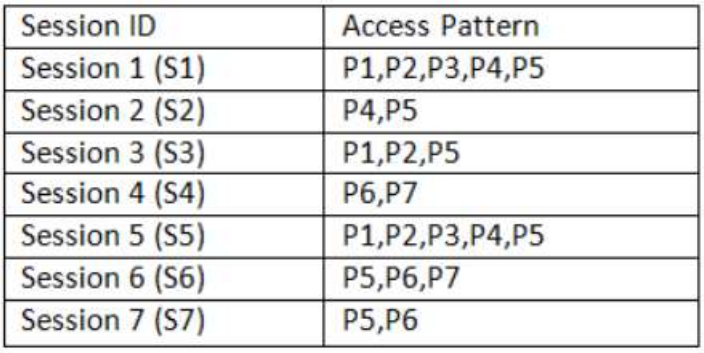
Consider an example that describes the whole flow and result of the system. It has some pages like google.com, yahoo.com, etc. Let's give code to those pages as shown in Table 4.1.

Table 4.1: Page Codes



Pages	Code
www.google.com	P1
www.yahoo.com	P2
www.facebook.com	P3
www.youtube.com	P4
www.ndtv.com	P5
www.gaana.com	P6
www.twitter.com	P7

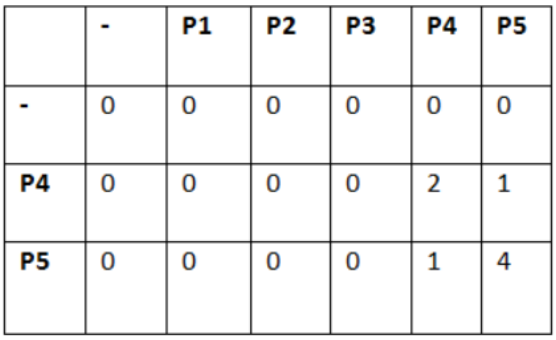
Table 4.2: Session Patterns



Session ID	Access Pattern
Session 1 (S1)	P1,P2,P3,P4,P5
Session 2 (S2)	P4,P5
Session 3 (S3)	P1,P2,P5
Session 4 (S4)	P6,P7
Session 5 (S5)	P1,P2,P3,P4,P5
Session 6 (S6)	P5,P6,P7
Session 7 (S7)	P5,P6

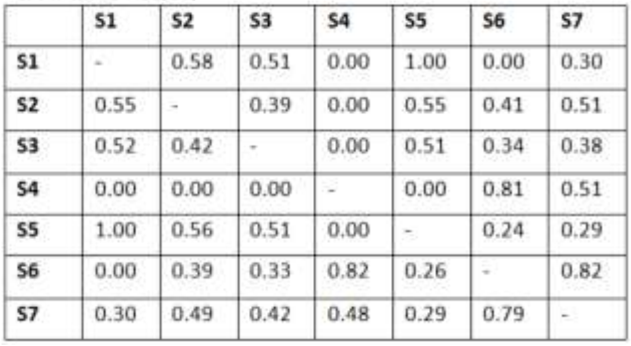
Now, consider session patterns as shown in Table 4.2. Initially, compute maximum similarity score between the sessions S1 and S2 using formula (1). For this, first find similarity score matrix of sessions S1 and S2 as shown in Table 4.3. Then find maximum value in this similarity matrix and maximum similarity value can be computed using formula (2). By following above sequence for all session patterns given in Table 4.2, maximum similarity score between all pairs of sessions as shown in Table 4.4.

Table 4.3: Similarity Score Matrix of Sessions S1 and S2



	-	P1	P2	P3	P4	P5
-	0	0	0	0	0	0
P4	0	0	0	0	2	1
P5	0	0	0	0	1	4

Table 4.4: Maximum Similarity Score



	S1	S2	S3	S4	S5	S6	S7
S1	-	0.58	0.51	0.00	1.00	0.00	0.30
S2	0.55	-	0.39	0.00	0.55	0.41	0.51
S3	0.52	0.42	-	0.00	0.51	0.34	0.38
S4	0.00	0.00	0.00	-	0.00	0.81	0.51
S5	1.00	0.56	0.51	0.00	-	0.24	0.29
S6	0.00	0.39	0.33	0.82	0.26	-	0.82
S7	0.30	0.49	0.42	0.48	0.29	0.79	-

Figure 4.1 shows the hierarchical clustering of similar sessions on x-axis and dissimilar on y-axis as the dendogram representation. After this, transition probability is calculated using formula (3) and output will be collected in the Transition Probability Matrix as shown in Table 4.5. It is described as, for an instance, probability of transition from page P1 to page P1 is 0.00, and similarly that of from page P1 to P2 is 1.00.

Figure 4.1: Dendogram representation

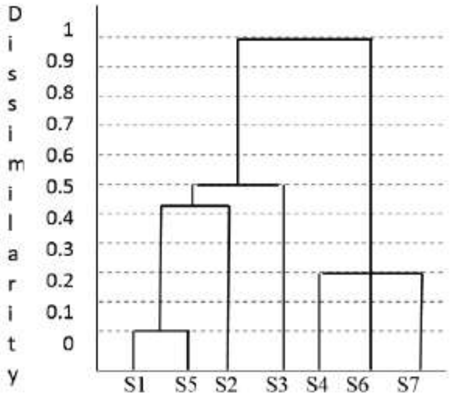
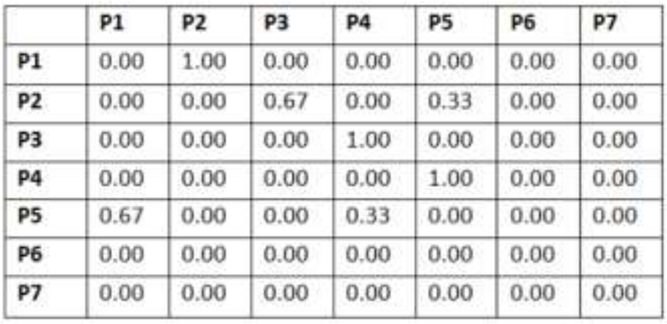


Table 4.5: Transition Probability Matrix



	P1	P2	P3	P4	P5	P6	P7
P1	0.00	1.00	0.00	0.00	0.00	0.00	0.00
P2	0.00	0.00	0.67	0.00	0.33	0.00	0.00
P3	0.00	0.00	0.00	1.00	0.00	0.00	0.00
P4	0.00	0.00	0.00	0.00	1.00	0.00	0.00
P5	0.67	0.00	0.00	0.33	0.00	0.00	0.00
P6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P7	0.00	0.00	0.00	0.00	0.00	0.00	0.00

V. Conclusion and Future Scope

Web usage mining is extraction of the information from web server log file which accessed by users. Web usage mining is done using three main steps pre-processing, pattern discovery and pattern analysis. Web Usage Mining is process of applying data mining techniques to the discovery of usage patterns from Web log data. In the proposed system, a new methodology using modified Levenshtein distance and higher order Markov model is developed to improve the web prediction accuracy. This system can be used to pre-fetch the web pages before they are actually being requested by the user, this reduces the access latency. Web page prediction is to predict the next web pages that user may access on internet based on the knowledge of previously visited web pages. This system avoids the irrelevant accesses by users and saves users surfing time.

As the designed system predicts the behavior of user by accessing the web logs and does the analysis on logs by applying the mining algorithms. As this technique works fluently but there can also be some advancement through which this system will be enhanced like building the system which concerns only for highly sensitive user, prediction of more than one upcoming sites which might be surf by user, categorizing type of users, and prediction of age of user.

References

[1]D. V. K. R. Harishkumar B T, Dr. Vibha L, “Web page access prediction using hierarchical clustering based on modified levenshtein distance and higher order markov model," 2016.

[2] R. P.Saravana kumar, R.Iswarya, “Predictive analysis of users behavior in web browsing and pattern discovery networks," International Journal of Latest Trends in Engineering and Technology, vol. 4, no. 1, pp. 239-245, May 2014.

[3] K. Z. Zhuang Like and Z. Changshui, “Session identification based on time interval in web log mining," Intelligent Information Processing II, pp. 389-396.

[4] G. V. P. Virendra R. Rathod, “Prediction of user behavior using web log in web usage mining," International Journal of Computer Applications, vol. 139, no. 8, p. 0975 8887, april 2016.

[5] S. G. J. R Geetharamani, P Revathy, “Prediction of users webpage access behavior using association rule mining," vol. 40, no. 8, p. 23532365, December 2015.

[6] D. A. J. P. G. Om Prakash, “Analyzing and predicting user behavior pattern from weblogs," International Journal of Applied Engineering Research, vol. 11, no. 9, pp. 6278-6283, 2016.

[7] R. K. S. Anurag kumar, Vaishali Ahirwar, “A study on prediction of user behavior based on web server log files in web usage mining," International Journal of Engineering And Computer Science, vol. 6, pp. 20 233-20 236, Feb 2017.

[8] D. R. K. K. R. Suneetha, “Identifying user behavior by analyzing web server access log file," International Journal of Computer Science and Network Security, vol. 9, no. 4, pp. 327-332, Apr 2009.

[9] Sana M. Deshmukh, K. P. Adhiya, “Understanding users navigation behavior using web mining algorithm," International Journal of Advanced Research in Computer and Communication Engineering, vol. 6, no. 8, pp. 214-221, Aug 2017.