IDENTIFY USER NAVIGATION PREDICTION USING WEB LOG DATA

[1]Narkhede Minal, [2]Mahajan Pramila, [3]Mutke Priyanka, [4]Vikas N Nirgude
Department of Computer Engineering S.R.E.S COE ,Kopargaon.
Email:[1]minalnarkhede677@gmail.com,[2]pramilamahajan4@gmail.com,
[3]mutkepriyanka@gmail.com, [4]vikas.nirgude2006@gmail.com

Abstract:
In order to improve web services the web mining includes the application of data mining techniques. Web Usage Mining (WUM) is the combination of both data mining and World Wide Web (WWW).Web usage mining mainly gives knowledge about the techniques that discover the user's usage pattern and try to predict the user's behaviors. The usage information about the user is recorded in logs of web server which are called as web log. Web log files are analyzed for the purpose of pattern extraction. This paper presents the Prediction of User navigation pattern using Clustering and Classification (PUCC) from web log data. In this paper we use k-means algorithm for the purpose of clustering and C4.5 algorithm to find potential user.

Keywords: WUM, PUCC, Web Mining, Clustering, Classification.

Introduction:

Web mining refers to the extraction of useful information and Knowledge from large amount of web data which can be used to improve web services. The important issue of web mining is to obtain useful information efficiently from large amount of web data [4].Web mining is divided into three areas: Web content mining (WCM), Web structure mining (WSM), and Web usage mining (WUM) .Web content mining mainly use for the retrieval of the useful information from the Web documents, while the Web structure mining use for the searching of how to model the underlying link structures of the Web[3] Web mining is very interested research topic which combines two of the activated research areas: World Wide Web and Data Mining. This paper mainly focuses on web usage mining. The process of extracting interesting patterns from the log data is known as web usage mining. Web usage mining mainly describes the techniques that discovers the user's usage pattern and try to predict the user's request[3]. Classification and clustering are two basic areas of machine learning research. Clustering divides a web log data into groups with similar interest. Classification is used to classify the data with similar category. Both areas are used for knowledge discovery[1].

System Architecture:

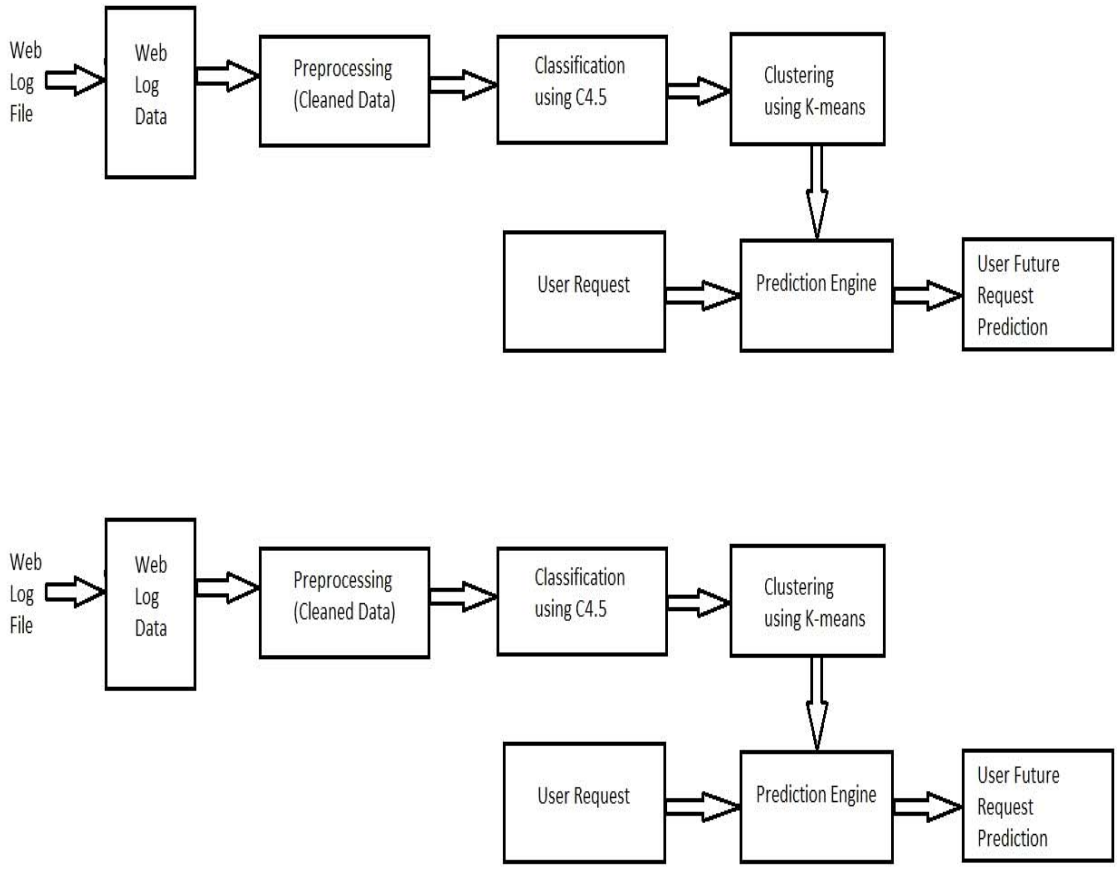

Fig.1 PUCC Model

Web log files: Web server automatically creates and maintains the Web log File. There are two formats of log File: Common Log File (CLF) and Extended Log File Format. Irrespective of the source of collection, the web log file has the following general characteristics:

1) The log file is in the form of text file in which records are identical in format.

2) A single HTTP request is represents by each record in the log file.

A log file record contains important information about a request: the client side host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL, and the browser information.A browser may fire multiple HTTP requests to Web server to display a single Web page. This is because a Web page not only needs the main HTML document; it may also need additional files, like images and JavaScript files. The main HTML document and additional files all require HTTP requests[1].

Pre-processing: Pre-processing of web log data is the first step of PUCC model. where the unformatted log data is converted into a form that can be directly applied to mining process. The

pre-processing steps include cleaning, user identification and session identification. Cleaning is the process in which all unwanted entries are removed.

Identification of Potential Users: This step is mainly focuses on separating the potential users from others. In this paper, we use C4.5 algorithm to find potential user.

Prediction Engine: Prediction Engine used to classify user navigation pattern and predicts user's future request.

Algorithms:

K-means:

Algorithmic steps for k-mean:

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$Vi = (1/Ci) \sum_{j=1}^{Ci} Xi \,(1)$$

Where, 'Ci' represents the number of data points in its cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3

C4.5 Algorithm:

Pseudo code :

A. Pseudo Code :

1. Check for base cases.
2. For each attribute a calculate: i. Normalized information gain from splitting on attribute a.
3. Select the best a, attribute that has highest information gain.
4. Create a decision node that splits on best of a, as root node.
5. Recurs on the sub lists obtained by splitting on best of a and add those nodes as children node.

Conclusion:

We are going to implement the PUCC model using K-means and C4.5 algorithms. In this system we perform preprocessing classification, & clustering. Then predict the users requests using prediction engine.

References:

[1].V. Sujatha, Punithavallib, "Improved user navigation pattern prediction technique from web log data",International Conference on Communication Technology and System Design 2011

[2] NeetuAnand et al, "Identifying the User Access Pattern in Web Log Data", / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (2), 2012, 3536-3539.

[3] Yan Wang, "Web Mining and Knowledge Discovery of Usage Patterns", CS 748T Project (Part I) February, 2000.

[4] Yue-She Lee, Show-Jane Yen, "Incremental and interactive mining of web traversal patterns", Department of computer Science and Engg,Mingchuan university,2007