International Conference On Emanations in Modern Technology and Engineering (ICEMTE-2017) ISSN: 2321-8169
Volume: 5 Issue: 3 221 - 223

Automated web pages data mining and recommendation system using K-NN classification method

Pallavi Kamalakar Bhoir and Prof. Neha Jain

Department of Computer Engineering
University of Mumbai

Shree L.R. Tiwari College of Engineering, Mira Road, Thane-401107

pallu.bhoir@gmail.com

Abstract **-** Web Mining can be classified into three main areas: Web Usage Mining, Web content Mining and Web Structure Mining. Web usage mining is a kind of web mining, which exploits data mining techniques to discover valuable information from navigation behavior of World Wide Web users. There are generally three tasks in Web Usage Mining: Preprocessing, Pattern analysis and Knowledge discovery. Preprocessing cleans log file of server by removing log entries such as error or failure and repeated request for the same URL from the same host etc... The main task of Pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users. The statistics collected from the log file can help to discover the knowledge. This knowledge collected can be used to take decision on various factors like Excellent, Medium, Weak users and Excellent, Medium and Weak web pages based on hit counts of the web page in the web site. The topology of the website is restructured based on user's behavior or hit counts which provides quick response to the web users, saves memory space of servers and thus reducing HTTP requests and bandwidth utilization. *The major problem of many on-line web sites is the presentation of many choices to the client at a time; this usually results to strenuous and time consuming task in finding the right product or information on the site.* In this work, an automatic web usage data mining and recommendation system based on current user behavior through his/her click stream data, in order to provide relevant information to the individual without explicitly asking for it. The K-Nearest-Neighbor (KNN) classification method will be used on-line to identify clients/visitors click stream data, matching it to a particular user group and recommend tailored browsing options that meet the need of the specific user at a particular time. To achieve this, web user's address file will be extracted, cleansed, formatted and grouped into meaningful session and data mart will be developed. The K-Nearest Neighbor classifier is transparent, consistent, straightforward, simple automated web usage data mining and recommendation system.

Index Terms **-** *Web mining, classification, Application, Tools, Algorithms, KNN algorithm.*

*****

I. INTRODUCTION

The Web mining is the process of using data mining techniques and algorithms to extract information from web side. Actually web mining is the application of data mining technique which is two types of data that is an unstructured or semi-structured data and it is automatically extract useful information or knowledge from web [1]. In web mining different application are website design, web search engines, information retrieval, network management, e-commerce, artificial intelligence and business. this application includes the temporal issues for the users.

Web mining can be classified into main tree areas: web usage mining, web content mining, and web structure mining. Each classification is having its own algorithms and tools. Web content mining is nothing but text mining; it is generally the second step in web data mining. Web content mining is the scanning and mining of text, hyperlink, pictures and graphs of a web page. Web structure mining it is one of the three categories of web mining for data.in web structure mining is a tool used to identify the relationship between web page. Web usage mining is also called as web log mining. Actually web usage mining data captures the identity or origin of web users and browsing behaviour at a web site [2].

Web mining process consists of four important steps: Resource finding, Data selection and pre-processing, Generalization and Analysis. In the resource finding process which is used to extract the data two form either online or offline text resources. In the Data selection and pre-processing is the process specific information from retrieved web and automatically selected and pre- processed. In the generalization used data mining and machine learning techniques to discover general patterns from the individual web sites and multiple web sides. During Analysis step using validation and interpretation of the patterns are done. Web mining can be classified into main tree areas: web usage mining, web content mining, and web structure mining [3].

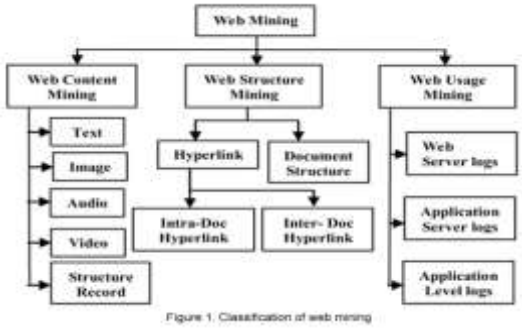


Figure 1. Classification of web mining

II. *RELATED WORK*

Some related works pertinent to this study, the review is specifically organized as follows:

A. Web data mining

Web data mining is an application of data mining techniques to discover patterns in web content, structure and usage. It is a branch of applied artificial intelligence that deals with storage, retrieval and

analysis of web log files in order to discover users accessing and usage pattern of web pages.

B. Forms of data mining system

Two forms of data mining tasks were identified by researchers over the years, these includes; predictive and descriptive. In predictive data mining task, inference is performed on current
data in a database in order to predict future values of interest while in descriptive task, data in a database are classified by characterizing the general properties of the data, it finds pattern describing the data in the database so as to present the interpretation to the user.

C. Classification of data mining system

Data mining system can be classified using different criteria. These criteria are identified as kind of database mined, kind of knowledge mined, and type of technique utilized and according to type of application adapted. In web usage data mining task, different techniques can be adopted, but the issue is how to determine which technique is most appropriate for the problem at hand. A multiple approach or an integrated technique that combines the benefits of a number of individual approaches can be adopted by a comprehensive data mining
System. There are different techniques for data classification which includes; decision tree classifier, Bayesian classifier, K-Nearest Neighbor classifier, and rule base classifier. In this
Work, the *K-Nearest Neighbor classification method* will be adopted.
Data mining techniques

D. *The K-Nearest Neighbor (KNN):*

A more scalable approach such as the KNN method, capable of handling training data that are too large to fit in memory is required. Many researchers have attempted to use K-Nearest Neighbor classifier for pattern recognition and classification in which a specific test tuple is compared with a set of training tuples that are similar to it. The theory of fuzzy set was introduced into K-Nearest Neighbor technique to develop a fuzzy version of the algorithm. The result of comparing the fuzzy version with the Crisp version shows that the fuzzy algorithm dominates its counterpart in terms of low error
Rate. In [1], The K-Nearest Neighbor algorithm was used alongside with five other classification methods to combine mining of web server logs and web contents for classifying users' navigation pattern and predict users' future request. The result shows that the KNN outperformed three of the other algorithms, while two of them performed uniformly. It was also observed that KNN archives the highest F-Score on the training set among the six algorithms. [2], as well adopted the KNN classifier to predict protein cellular localization site. The result of the test using stratified cross-validation shows the KNN classifier to perform better than the other methods which includes binary decision tree classifier and the naïve Bayesian classifiers.

II. MOTIVATION / Justification for using KNN algorithm over other existing algorithm

The K-Nearest Neighbor (K-NN) algorithm is one of the simplest methods for solving classification problems; it often yields competitive results and has significant advantages over
Several other data mining methods. This work will therefore be based on the need to establish a flexible, transparent, consistent straightforward, simple to understand and easy to implement approach. This will be achieved through the application of K-Nearest Neighbor technique, which will be able to overcome some of the problems associated with other
Available algorithms. It is able to achieve these by the following:

– Overcoming scalability problem common to many existing data mining methods such as decision tree technique, through its capability in handling training data that are too large to fit in memory.

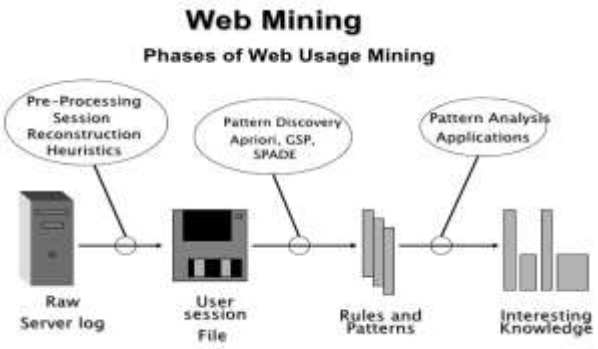
– The use of simple Euclidean distance to measure the similarities between training tuples and the test tuples in the absence of prior knowledge about distribution of data therefore makes its implementation easy.

– Reducing error rate caused by inaccuracy in assumptions made for usage of other technique such as the Naive Bayesian classification technique, such as class conditional independency and the lack of available probability data which is usually not the case when using KNN method.

– Providing a faster and more accurate recommendation to the client with desirable qualities as a result of straightforward application of similarity or distance for the purpose of Classification.

III. Overview of steps in performing web usage data mining task

Data mining task can be categorized into different stages based on the objective of the individual analyzing the data. The overview of the task for each step is presented in detail as follows:



A. Data acquisition, preprocessing and data mart development

Data acquisition:
This refers to the collection of data for mining purpose, and this is usually the first task in web mining application [3]. The said data can be collected from three main sources which includes (i) web server (ii) proxy server and (iii) web client [4]. In this work, the web server source will be chosen for the fact that it is the richest and most common data source; more so, it can be used to collect large amount of information from the log files and databases they represent. The user profile information, the access and navigation pattern or model is extracted from the historical access data recorded in the reader site, users' address database. The data are so voluminous as it contains so many detailed information such as date, time in which activities occur, saver's name, IP address, user name, password, dailies name, required feed, news headlines, and contents, as recorded in the database file.

B. Data pre-processing:

In the original database file extracted, not all the information is
Valid for web usage data mining, it needs only entries that contain relevant information. The original file will usually be made up of text files that contain large volume of information
Concerning queries made to the web server in which in most instances contains irrelevant, incomplete and misleading information for mining purpose. Data preprocessing is described as the cleansing, formatting and grouping of web log files into meaningful session for the sole aim of utilizing it for web usage mining.

C. Data cleansing:
Data cleansing is the stage in which irrelevant/noisy entries are
Eliminated from the log file [5]. For this work the following operations will be carried out: (i) Removal of entries with ''Error'' or ''Failure'' status. (ii) Removal of requests executed by automated programs such as some access records that are automatically generated by the Search engine agent from access log file and proxies. (iii) Identification and removal of request for picture files associated with request for a page and request include Java scripts (.js), and style sheet file (iv) Removal of entries with unsuccessful HTTP status code, etc.

D. Data mart development:
Data mart is a logical subset of data warehouse. If the data warehouse DBMS can support more resources, that will be required of the data mining operation, otherwise a separate data mining database will be required. Since the raw log file is usually not a good starting point for data mining operation, the development of a data mart of log data is required for the data mining operation. In this work a separate data mart of users' address URL will be developed using relational database Management software
MySQL.

E. Transaction identification
There is need for a mechanism to distinguish different users so as to analyze user's access behavior [4]. Transaction identification is meant to create meaningful clusters of references for each user. A user navigation behavior can be represented as a series of click operations by the user in time sequence, usually call click stream, which can further be divided into units of click descriptions usually referred to as session or visit.

F. Session identification:
A session can be described as a group of activities carried out by a user from the user's entrance into the web site up to the time the user left the site. It is a collection of user clicks to
a single web server [6]. Session identification is the process of partitioning the log entries into sessions after data cleansing operation. Here, the cookies are used to identify individual users, so as to get a series of clicks within a time interval for an identified user. One session can be made up of two clicks, if the time interval between them is less than a specific period [4], [5].

G. Pattern discovery
Pattern discovery is the key process of web mining which includes grouping of users based on similarities in their profile and search behavior. There are different web usage data mining techniques and algorithms that can be adopted for pattern discovery and recommendation, which includes path analysis, clustering, and associate rule. In this work, *the K-Nearest Neighbor classification technique* will be used in order to observe and analyze user behavior pattern and click stream from the pre-*process* to web log stage and to recommend a unique set of object that satisfies the need of an active user, based on the users' current click stream.

H. Pattern analysis
Pattern analysis is the final stage in web usage mining which is aimed at extracting interesting rules, pattern or statistics from the result of pattern discovery phase, by eliminating irrelevant rules or statistics. The pattern analysis stage provides the tool for the transformation of information into knowledge. We have incorporated an SQL language to develop a data mart using MySQL DBMS software specifically created for web usage mining purpose in order to store the result of our work [7]. The data mart will be populated from raw users address URL file of the reader's site that contains some basic fields needed.

IV. The working of K-Nearest Neighbor classifier

The K-Nearest Neighbor classifier usually applies either the Euclidean distance or the cosine similarity between the training tuples and the test tuple but, for the purpose of this research work, the Euclidean distance approach will be applied in implementing the K-NN model for our recommendation system [8]. In this work, suppose our data tuples are restricted to a user or visitor/ client described by the attribute Daily Name, Daily Type and News category and that X is a client with Day as username and Dy123 as password. The Euclidean distance between a training tuple and a test tuple can be derived as follows:

Let Xi be an input tuple with p features (xi1,xi2, . . .,xip)
Let n be the total number of input tuples (i =1,2,. . .,n)
Let p be the total number of features (j =1,2,. . .,p)

The Euclidean distance between Tuple can be calculated. In K-NN, classification, all neighboring points that are nearest to the test tuple are encapsulated and recommendation will be made based on the closest distance to the test tuple, The nearest tuple is determined by the closest distance to the test tuple. The K-NN rule is to assign to a test tuple the majority category label of its K-Nearest training tuple.

VI. *CONCLUSION AND FUTURE WORK*

In this paper we discussed about the basic automatic Real-Time recommendation system. The system perform classification of users on the simulated active sessions extracted from testing sessions by collecting active users click stream and matches this with similar class in the data mart .
The K-NN classification model implemented with Euclidean distance method is capable of producing useful and a quite good and accurate classification to the client.

REFERENCES

[1] L. Habin, K. Vlado, Combining mining of web server logs and web content for classifying users' navigation pattern and predicting users future request, J. Data Knowledge Eng. 61 (2007) (2006) 304–330, http://dx.doi.org/10.1016/j.datak.2006.06.001.
[2] H. Paul, N. Kenta, Better Prediction of Protein Cellular Localization Sites with the Nearest Neighbor Classifier, ISMB-97, Proceeding of America Association for Artificial Intelligence, USA, 1997, pp. 147–152.
[3] A. Dario, B. Eleno, B. Giulia, C. Tania, C. Silvia, M. Naeem, Analysis of diabetic patients through their examination history, J. Expert Syst. Appl. 40 (2013) 4672–4678, http://dx.doi.org/10.1016/ j.eswa.2013.02.006.
[4] M.F. Federico, L.L. Pier, Mining interesting knowledge from weblog: a survey, J. Data Knowledge Eng. 53 (2005) (2005) 225–241, http://dx.doi.org/10.1016/j.datak.2004.08.001.
[5] M. Michal, K. Jozef, S. Peter, Data preprocessing evaluation for web log mining: reconstruction of activities of a web visitor, J. Proc. Comput. Sci. 1 (2012) (2012) 2273– 2280, http://dx.doi.org/10.1016/j.procs. 2010.04.255.
[6] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining World Wide Web browsing patterns, J. Knowledge Inform. Syst. 1 (1) (1999) 1–27.
[7] T. Luigi, S. Giacomo, Mining frequent item sets in data streams within a time horizon, J. Data Knowledge Eng. 89 (2014) 21–37, http://dx.doi.org/10.1016/j.datak.2013.10.002.
[8] H. Jiawei, K. Micheline, Data mining concept and Techniques, second ed., Morgan Kaufmann Publishers, Elsevier inc., USA San Francisco, CA 94111, 2006, p. 285–350.