ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Data Mining of Web Access Logs Using Classification Techniques

Md. Azam[1], Asst. Prof. Md. Tabrez Nafis[2]

[1]*M.Tech Scholar, Department of Computer Science & Engineering, Al-Falah School of Engineering & Technology, Dhauj, Faridabad, Haryana, INDIA*

[2]*Asst. Prof. Department of Computer Science Faculty of Management & IT, Jamia Hamdard,*

***Abstract:** This paper focus on classification, technological advancement, business have gone online. Because of the growing popularity of the World Wide Web. Many website typically experience Lac of visitors every day. Data mining techniques such as association rules, classification, and clustering and attribute selection are considered very useful in web usage mining. Data Mining is gaining more popularity because of its power to extract knowledge from voluminous data where it is beyond the reach of traditional techniques of knowledge discovery and human comprehension. A data mining tool, WEKA is used. The WEKA is a collection of machine learning algorithms for solving real-world data mining problems.*

***Keyword:** Data Mining, Data Mining Techniques, Data Classification.*

I. INTRODUCTION

Data Mining is ahead more recognition because of its power to extract knowledge from voluminous data where it is beyond the reach of traditional techniques of knowledge discovery and human understanding. The entire process of data mining can be divided into 4 phases: data collection, data preparation, pattern discovery and pattern analysis.

The data collection phase involves collection of data from different sources and identifying desired features for mining. Data collected from different sources may be heterogeneous. The data preparation phase involves the process of normalizing the data and representing them in a structure so that they become more convenient. Features recognized in the previous phase are extracted and at last, formatting is applied to represent data in a format required by the data mining tool to be used. Mining Techniques

Data mining techniques can be used to find access patterns hidden inside huge volumes of web access data. Once the data are prepared and formatted, the data mining techniques are applied to discover patterns. For the experiments of this research paper, a data mining tool, WEKA is used. The WEKA is a collection of machine learning algorithms for solving real-world data mining problems. Data mining tools for preprocessing, classification, clustering, attribute selection and visualization are implemented in the WEKA.

Prototype Analysis

Prototype analysis is the final phase of the knowledge discovery process. In this phase, the discovered patterns are judged for their interestingness. The exact methodology depends upon the application for which mining is done [3]. The analysis can be done using the content and structure of the website. Visualization techniques that represent the values at different positions using different colors are often useful in the Prototype analysis [3].

Network Mining

Incredible growth of the World Wide Web, the raw web data has become a gigantic source of information. accordingly, this has turned researchers attention towards the use of data mining

ISSN: 2321-9653

techniques to this data. Network Mining is referred to as the application of Network mining technologies to the web data [4].

Network mining can be categorized into content mining, structure mining or usage mining depending upon which part of the web to mine [5].

Content Mining

The web content is the authentic data the web page was deliberate to convey to the users. It consists of several types of data such as unstructured text, graphics, sound, video and semi-structured hypertext. Content mining can be referred to as the application of data mining algorithms to the content of the web [3]. A conceptual schema can be created [6] that can describe the semantics of a large volume of unstructured web data to manage them [5]. Discussed various categories of the web content mining such as text mining which is mining of unstructured texts and multimedia data mining which is mining of multiple types of data such as unstructured and image data.

Net Access Logs and Net Usage Mining

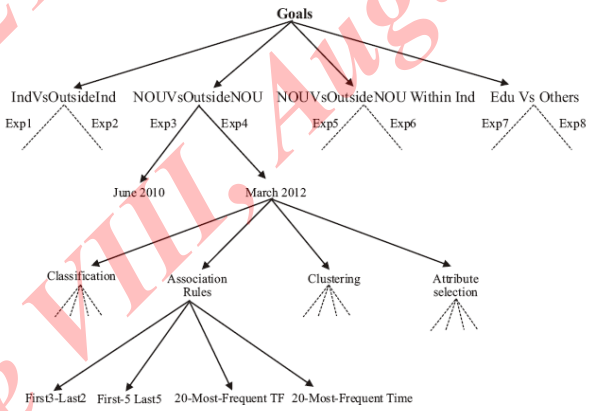
The Net access log is the main resource for Net usage mining because it stores data pertaining to accesses of the website. The usage data can be stored in Common Log Format (CLF) or Extended Log Format (ELF). The Net access log in CLF format has information of the IP address of a visitor's machine, the user id of visitor if available and date/time of the page request. The method is a means of page request. It can be GET, PUT, POST or HEAD. The URL is the page that is requested. The protocol is the means of communication used, HTTP for example. The status is the completion code. For example, 500 is the code for success. The size field shows the bytes transferred as a result of a page request. The Extended Log Format, in addition to these information, stores referrer, which is the page this request has come from and agent is the web browser used.

Data Mining Experiments

This section brief describes the organization of experiments conducted and the pattern discovery tasks performed for each experiment. A hierarchical view of the pattern discovery tasks performed in the experiments of this paper and is intended to show the organization of the experiments.

The access log files of the June-2012 and the March-2014 were used for the experiments. One experiment addresses one goal on one access log file. For example, experiment was conducted for the goal Nalanda Open Univeristy VsNot Nalanda Open University using the March-2014 log file.

For each experiment, data mining techniques were used. These techniques were classification, association rules, clustering and attribute selection. A pattern discovery task for one of these data mining techniques was performed using one of four different feature sets extracted from the preprocessed data. The four different feature sets used for the experiments are First3-Last2, First5-Last5, 20-Most-Frequent-TF and 20-Most-Frequent-Time. As an example, a pattern discovery task can be performed to find association rules using the First3-Last2 feature set.



A Hierarchical View of the Pattern Discovery Tasks Performed

Web Log Data

Information about web log data used for the experiments. The web crawler entries were not useful for the experiments. crawlers generally access “robots.txt” file for the website access permissions. Therefore, these entries were removed by preparing a list of crawlers that accessed the “robots.txt” file. However, this technique does not remove all crawler entries.

The log entries of IP addresses that no longer existed were also not useful for the experiments. Hence, those entries were removed.

ISSN: 2321-9653

The image entries, entries by proxy servers and entries of bad requests, were not useful for the experiments of this paper. Therefore, they were removed.

Web Access Logs File	Number of Entries	Time Period
Access 2010	1000000	29/05/2010-03/06/2010
Access 2012	11391157	04/02/2012 – 23/04/2012

Detail of Web Access Log Files

Operational Identification

An IP-Day consists of all entries of the pages visited from one IP address in a day. Once the extraneous entries were removed from the web log data, IP-Days were extracted based on the IP addresses. The transactions were identified from the IP-Days by the time duration between two consecutive visits. If the time duration between accesses of two pages "X" and "Y" in an IP-Day is more than 30 minutes then "X" was considered as the last page accessed in one transaction and "Y" was considered as the first page accessed in another transaction. A time period of 30 minutes was considered appropriate to distinguish two transactions. It may be possible that a visitor starts a visit at 12:55 pm and ends at 00:11 am in which case there will be two transactions generated instead of one.

It was conjectured that transactions that have at least 5 pages visited would be useful for this data mining task. Transactions with fewer than 5 visited pages were not used. The number of transactions used for the experiments from each log file.

Web Access Log file	Number of Transaction
Access 2010	4591
Access 2012	55602

Number of Transaction Used from Web Access Log Files

Experiment: IndVsOutsideInd2012

The experiment was conducted to compare access patterns between visitors from within India and visitors from outside India using the web access log file "access2012".

Classification

The results were analysed only if the classification accuracies were above 70 %.

20-Most-Frequent-TF

The output of the J48 and the 1R algorithms indicated the root as the most significant attribute. The partial decision tree that if visitors visit the root page at all then they are from within India. This result is consistent with the result obtained using the 20-Most-Frequent-TF feature set.

Data Mining Technique	Web Access Log File	Feature Set Used	Significant Pattern Discovered
Classification	Access 2012	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES
Association Rules	Access 2012	First3-Last2	NO
		First5-Last5	YES
		20-Most-Frequent-TF	NO
Clustering	Access 2012	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	NO
		20-Most-Frequent-Time	NO

ISSN: 2321-9653

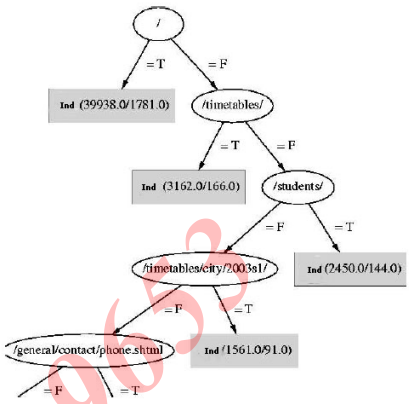
Attribute Selection	Access 2012	First3-Last2	YES
		First5-Last5	YES
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES

Experiment2: IndVsOutsideInd2012 - Summary of results

The decision tree also shows that if visitors do not visit the root page and visit the "/timetables" page then they are from within India. This may be because visitors from within India tend to look at the timetables of various subjects of their interest. It may be possible that this result is dominated by students of Nalanda Open University who know the URL of the "/timetables" page and hence visit this page directly. The decision tree also shows that if visitors do not visit the root page and they also do not visit the "/timetables" page but visit "/students" page then they are from within India. This result is obtained may be because the students of Nalanda Open University visit this page frequently for various activities such as checking the emails.

20-Most-Frequent-Time

The output of the 1R algorithm indicated "/courses" as the most significant attribute. The partial decision tree the root as the most significant attribute. The decision tree also shows that if visitors do not visit the root page and visit the "/timetables" page then they are from within India. The decision tree also showed that if visitors do not visit the root page and they also do not visit the "/timetables" page but they visit "/students" page then they are from within India. This outcome is consistent with the result obtained using the 20-Most- Frequent-TF feature set.



Partial Decision Tree of the WEKA J48 Program Using the 20-Most-Frequent-TF Feature Set in Experiment

Association Rules

The association rules generated by the Apriori algorithm were analysed to find their interestingness.

First5-Last5

The Apriori algorithm found one interesting rule which can be interpreted as if visitors visit the "/timetable" page then they are from within India. This result is consistent with the results obtained using the 20-Most-Frequent-TF feature set.

Attribute Selection

The attributes selected by the process of Attribute Selection using "cfssetEval" attribute evaluator together with "BestFirst" search method were analysed for their interestingness.

First3-Last2

Attribute selection using this feature set indicated "link0", which is the first page visited in a transaction, as the most significant attribute. This result is consistent with the results obtained the First3-Last2 feature set.

First5-Last5

Attribute selection using this feature set indicated “link0” as the most significant attribute. This result is consistent with the result obtained using the First3-Last2 feature set.

20-Most-Frequent-TF

Attribute selection using this feature set indicated the root page as the most significant attribute. This result is consistent with the result using the 20-Most-Frequent-TF feature set.

20-Most-Frequent-Time

Attribute selection using this feature set indicated the root page, “/students" and “/timetables" as the most significant attributes in order. This result is consistent with the result shown by decision trees using the 20-Most-Frequent-TF and the 20- Most-Frequent-Time feature sets.

II. CONCLUSIONS

The goal of the experiments was to compare access patterns between visitors from within India and visitors from outside India. Visitors from within India mostly visit the root page, whereas visitors from outside India mostly visit specific pages directly. This is probably because they use search engines.

REFERENCES

[1] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations, 1(2):12{23, 2000.

[2] J. Borges and M. Levene. Data mining of user navigation patterns. In Proc. of the Workshop on Web Usage Analysis and User Profiling (WEBKDD'99), pages 92{111, 2001.

[3] R. Kosala and H. Blockeel. Web mining research: A survey. ACM SIGKDD, 2(1):1{15, 2000.

[4] K.W. Tan, H. Han, and R. Elmasri. Web data cleansing and preparation for ontology extraction using WordNet. In First International Conference on Web Information Systems Engineering (WISE'00), volume 2.

[5] B. Mobasher, R. Cooley, and J. Srivastava. Creating adaptive web sites through usage-based clustering of urls. In IEEE Knowledge and Data Engineering Workshop (KDEX'99), 1999.

[6] C. R. Anderson. A machine learning approach to web personalization. PHD Thesis, University of Washington, 2002.