International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

WEB PAGE PREDICTION USING WEB MINING

B. RAJESWARI[1], Dr. S. SHAJUN NISHA[2]

[1] *Research Scholar, Department of Computer Science, Sadakathullah Appa College, Tiruneveli, Tamil Nadu, India*
[2] *Professor, Department of Computer Science, Sadakathullah Appa College, Tiruneveli, Tamil Nadu, India*

Abstract - *Web mining is an obvious and popular one of the data mining techniques. Web mining is defined as the use of data mining techniques to automatically discover and extract information from the web document and services. When the web user access the network a large amount of data is generated and it is stored in web log files. The web log file contains previous user navigation data or historical data. This web access pattern is used to find the user access behavior. Through this behavior analysis it is much easier to predict the next set of pages the user going to visit. This approach is based on session, Time and frequency based analysis using comparison of different classification techniques such as c4.5, AdaBoost M1, SVM, Rule Part. Finally predict the accuracy result for web page recommendations.*

Key Words: web usage mining, k-means clustering, c4.5, AdaBoost M1, SVM, Rule Part.

I.INTRODUCTION

The World Wide Web is the explosive growth of knowledge available in the internet. *Data mining* is defined as a process used to extract usable data from a larger data set, Web mining is the one of the techniques of data mining process. Web mining is needed for predicting the user access behavior to improve the usability and user access maintenance of the web sites. Web page recommendation plays a vital role in web page systems. Web Page recommendations system are used to implemented on Web server and make use of data obtained as a results of the different collections of user browsing patterns and explicit data. Useful knowledge discovery from web access log, which lacks an integrated structure or schema, it becomes much more difficult for users to access relevant information efficiently. The user browsing data will be processed via the interaction of users and web site depend on the web site itself, for example some web site require user name, password, email address, web site collects past purchases or visit histories of users, as well as other explicit and implicit representation of users' interests reviews, these data can be used to represent users' profile which allows the recommendation system to group of similar users and recommended for new web pages. Web usage mining includes the data processing method which is consisting of converting the usage, content and structure information. Select the web log files which are already stored in the system server. Web log file is processed to remove the unwanted data from the log files. Find out the targeted attributes and evaluate the similar clustering data by using k-means clustering algorithm. K-means clustering algorithm is using k values and find similar data can be grouping into a clustering mechanism. Analysis the user behavior pattern can be divided into three key domain Time, Frequency, and Session domain analysis. Time based analysis is used to evaluate the amount of time spent on a particular URL in the entire web log files. Calculate the frequency based analysis is used to evaluate the user navigational frequency count of each URL. Finally calculate the session based analysis is used to evaluate the morning, evening and afternoon accessed session of an URL in the entire web log files. Then apply different types of classification method such as c4.5, Ada Boost M1, Support Vector Machine and Rule Part classifier algorithms. The C4.5 algorithm is used to converting trees to rules and handle with missing value. The classification performance of c4.5 and SVM classifier are considered more reliable. Ada Boost M1 perform very well for data sets. The different classification methods are used to find out the accuracy results for new web page recommendations.

1.1 RELATED WORK:

In this paper describe web usage mining can be divided into three key domain Time, Frequency and Session domain [1].

This work focus on different types of classification method such as c4.5, Ada Boost M1, Support Vector Machine and Rule Part classifier algorithm. The C4.5 algorithm is used to converting trees to rules and handles the missing value. The classification performance of c4.5 and SVM classifier are considered more reliable [2].

Time based analysis is used to evaluate the amount of time spent on a particular URL in the entire web log files. Calculate the frequency based analysis is used to evaluate the user navigational frequency count of each URL. Finally calculate the session based analysis is used to evaluate the morning, evening and afternoon accessed session of URL in the entire web log files [3].

AdaBoost M1 perform very well for data sets with two classes. But it is too restrictive to data sets with more than two classes simple and weak. AdaBoost is an algorithm for constructing a "strong" classifier as linear combination f(x) = XT t=1 αtht(x) of "simple" "weak" classifiers ht(x)[4].

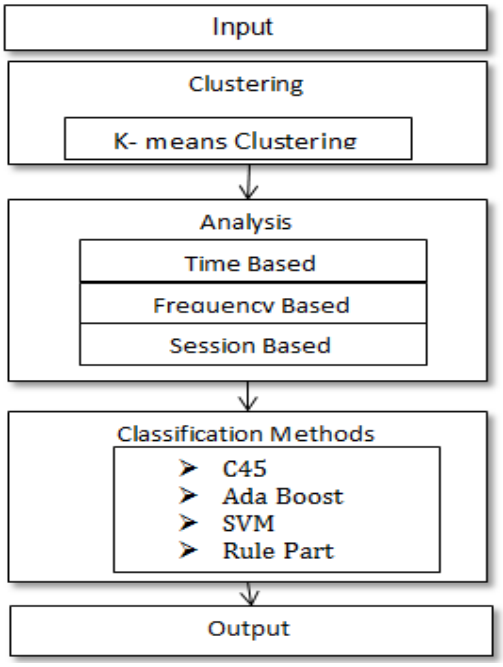
The C4.5 algorithm is belongs to the decision tree algorithm group. This algorithm takes input in the form of testing URLs. Testing dataset in the form of sample URL dataset that will be used to build a tree that has been substantiated. While samples are data field that will be used as a parameter within the classification data. C4.5 algorithms are algorithms result of the development of the algorithm ID3. Improvements from ID3 algorithm C4.5 algorithms performed in the case (Santosa, 2003) [5].

Support vector machines select a small number of critical boundary instance called support vectors from each class and build a linear discriminant function that separates them as widely as possible. Molecules in the test are mapped to the same future space and their acitivity is predicted according to which side of the hyper plane they fall. The distance to the boundary can be used to assign confidence level to the prediction such that higher the distance the higher the confidence [6].

1.2. MOTIVATION AND JUSTIFICATIONS:

Web mining has several advantages over current methods of collecting information like the use of cookies and cache. K-means clustering method is used to measurable and efficient in large data collection. C4.5 algorithm is ability to use attributes with different weights. AdaBoost classifier much is easier to implement. Support Vector Machine (SVM) based on dense concept to support the high dimensional input space. Support Vector Machine (SVM) based on dense concept to support the high dimensional data. So the work is to justify the Rule Part classifier is easy to generate rules and reduce problem complexity

1.3 OUTLINE OF THE PROPOSED WORK



Method:

- Take an web page URLs as input data set
- Pre-processing the Dataset
- Time Based Analysis: It is based on time duration
- Frequency Based Analysis: It is based on user pattern and navigational pattern
- Session Based Analysis: It is based on session
- Different Classify using c4.5, Ada Boost, SVM, Rule Part Classifier
- Compare the different classifier techniques
- Finally find the Accuracy output for dataset

1.4. ORGANIZATION OF THE PAPER

This paper is organized as follows: In Section II describes the K-means clustering and Different types of classification methods. Section III Display the Experimental results, and conclusion is placed in section IV

II. METHODOLOGY

In this paper different types of classification methodology and clustering concept are used to find the accuracy of web page prediction.

2.1 DATASET

There are 100 entries and 48 instances. Take an web page URLs as input data set

2.2 CLUSTERING

2.2.1 K-MEANS CLUSTERING

K-means clustering algorithm was respectively performed on pre-processing data. K-means clustering using k values ranging from 1 to 50 as the input preprocessed urls.

2.3 Classifications

2.3. 1 C4.5

The C4.5 algorithm is belongs to the decision tree algorithm group. This algorithm has input in the form of testing URLs. Testing dataset in the form of sample URL dataset that will be used to build a tree that has been substantiated. While samples are data field that will be used as a parameter within the classification data. C4.5 algorithms are algorithms result of the development of the algorithm ID3. Improvements from ID3 algorithm C4.5 algorithms performed in the case (Santosa, 2003):

1. C4.5 can handle with missing value
2. Deal with continuous data
3. Pruning
4. Convert trees to rules

2.3.2 ADABOOST

AdaBoost M1 perform very well for data sets with two classes. But it is too restrictive to data sets with more than two classes simple and weak. AdaBoost is an algorithm for constructing a ”strong” classifier as linear combination f(x) = X T t=1 αtht(x) of “simple” “weak” classifiers ht(x).

Strength of ada boost algorithm is fast, simple and easy to program.

AdaBoost.M1 performs very well for data sets with only two classes, but it is too restrictive to data sets with more than two classes. AdaBoost.M1 performs very well for data sets with only two classes, but it is too restrictive to data sets with more than two classes.

- ht(x) ... “weak” or basis classifier, hypothesis, ”feature”
- H(x) = sign(f(x)) ... ‘‘strong” or final classifier/hypothes

2.3.3. SVM (Support Vector Machine)

Support vector machines select a small number of critical boundary instance called support vectors from each class and build a linear discriminant function that separates them as widely as possible. Molecules in the test are mapped to the same future space and their acitivity is predicted according to which side of the hyper plane they fall. The distance to the boundary can be used to assign confidence level to the prediction such that higher the distance the higher the confidence

2.3.4 Rule Part

Many classification rule algorithms are used to generate the classification rules such as ID3, CART, and uRule. ... Keywords Data Mining, Classification Rules, C4.5, Ripper, Part I.

III. EXPERIMENTAL RESULT

Suppose a user accessed the following ten pages. Table 1 contains the Example URL data

Table 1: Example URLs

S.No.	URLs
1	www.facebook.com
2	www.google.com
3	www.twitter.com
4	www.msn.com
5	www.rediffmail.com
6	www.gmail.com
7	www.yahoo.com
8	www.shopclues.com
9	www.amazon.com
10	www.flipkart.com

Table 1 contains the Example URL data

Table 2 : URL Frequency

S.No.	Websites	Frequency
1	www.facebook.com	12
2	www.google.com	14
3	www.twitter.com	8
4	www.msn.com	6
5	www.rediffmail.com	7
6	www.gmail.com	10
7	www.yahoo.com	12
8	www.shopclues.com	18
9	www.amazon.com	22
10	www.flipkart.com	24

Table 2 contains the URLs for a single user and the amount of times the URL navigated by the target user.

Table 3 : Time Computation

SNo	Websites	Time(In Seconds)
1	www.facebook.com	3256
2	www.google.com	7093
3	www.twitter.com	7156
4	www.msn.com	7196
5	www.rediffmail.com	6889
6	www.gmail.com	7156
7	www.yahoo.com	7196
8	www.shopclues.com	6889
9	www.amazon.com	10516
10	www.flipkart.com	10696

The table 3 contains the URLs for a single user and the time consumed by the target user

Table 4 : URL Observation Session

SNo	Websites	S1	S2	S3
1	www.facebook.com	16	11	24
2	www.google.com	17	19	23
3	www.twitter.com	18	27	22
4	www.msn.com	7	3	10
5	www.rediffmail.com	11	7	16
6	www.gmail.com	21	17	26
7	www.yahoo.com	12	8	15
8	www.shopclues.com	9	6	14
9	www.amazon.com	14	9	21
10	www.flipkart.com	13	10	18

The above Table 4 contains the data from a single user is sub-divided into three observations.

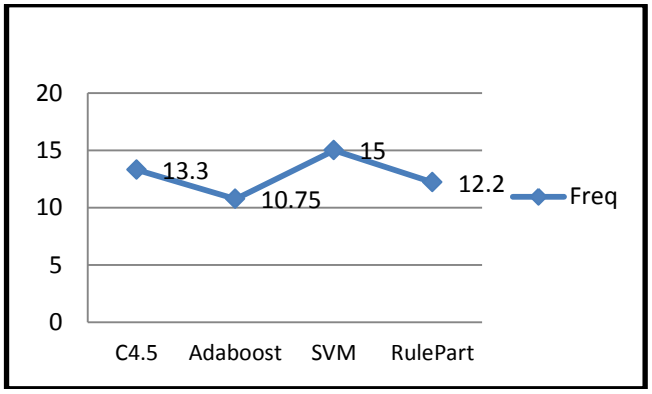


Figure1: Frequency Analysis

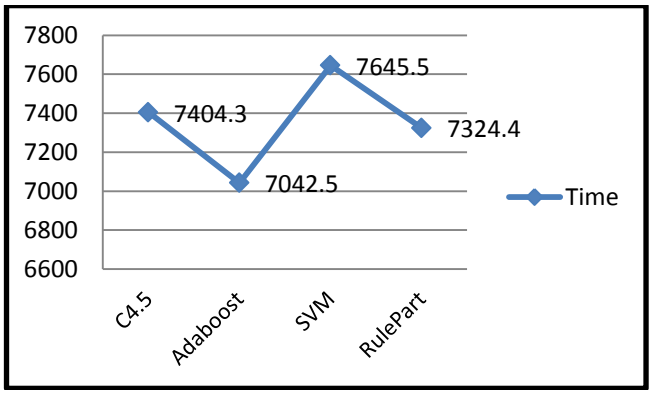


Figure2 : Time Analysis

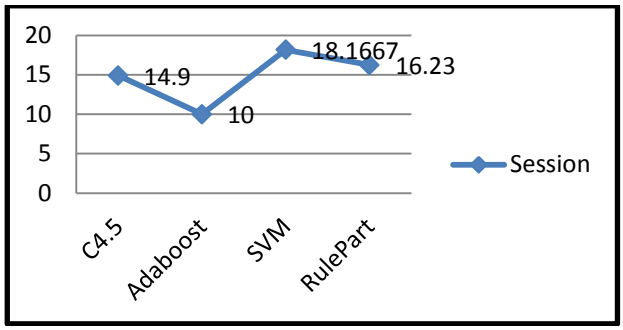


Figure3 : Session Analysis

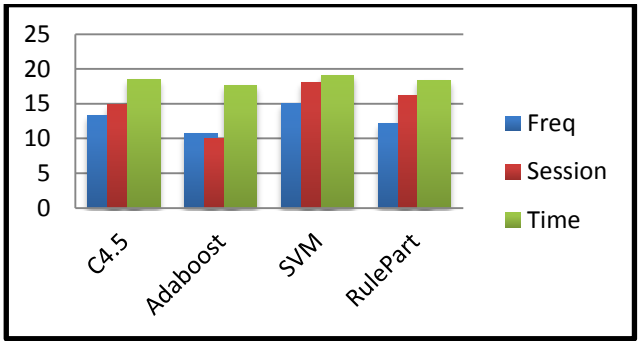


Figure4 : Output Result

IV. CONCLUSION

An intelligent web-page recommendation system based on frequency, Time and session. In the proposed system, different classification algorithms are used to predict the accuracy results. The proposed system has achieved good performance with high satisfaction and applicability and the time required for predicting the next web-pages using best classifiers.

IV. REFERENCES

[1] B.Rajeswari, Dr.S.Shajun Nisha, "Web Page Recommendation using Web Usage Mining Based on Session Based Analysis" Sadakath: A research Bulletin, Volume .V, Feb.2018

[2] Haibin Liu, Vlado Keseli "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests" Data and Knowledge Engineering, Volume 61,Issue 2, May 2007

[3] Sheetal Kumrawat, Pramod S. Nair, "Web Page Recommendation Using Efficient Weight Based Prediction System", International Journal Of Science and Research(IJSR), Volume 4 Issue 11, November 2015

[4] www.google.com

[5] www.slideshare.com