

WEB DATA USING MINING TECHNIQUES WITH THE HELP OF OPEN SOURCE SOFTWARE PACKAGES

Harsh Mathur

ABSTRACT

Predicting the next page to be accessed by the Web users has attracted a large amount of research. In this paper, a new web usage mining approach is proposed to predict next page access. It is proposed to identify similar access patterns from web log using K-mean clustering and then Markov model is used for prediction for next page accesses. The tightness of clusters is improved by setting similarity threshold while forming clusters. In traditional recommendation models, clustering by non-sequential data decreases recommendation accuracy. In this paper involve incorporating clustering with low order markov model which can improve the prediction accuracy. The main area of research in this paper is pre processing and identification of useful patterns from web data using mining techniques with the help of open source software packages.

KEYWORDS: Web Usage Mining, Rough Set Model, Conditional Attributes

1. INTRODUCTION

Rough set theory can be regarded as a new mathematical tool for imperfect data analysis. The theory has found applications in many domains, such as decision support, engineering, environment, banking, medicine and others. This paper presents basis of the theory which will be illustrated by a simple example of churn modeling in telecommunications. Rough set philosophy is founded on the assumption that with every object of the universe of discourse some information (data, knowledge) is associated. Objects characterized by the same information are *indiscernible (similar)* in view of the available information about them. The *in-discernibility relation* generated in this way is the mathematical basis of rough set theory. Any set of all indiscernible (similar) objects is called an elementary set, and forms a basic granule (*atom*) of knowledge about the universe. Any union of some *elementary sets* is referred to as a crisp (*precise*) set – otherwise the set is rough (*impre-cise, vague*). Each rough set has boundary-line cases, i.e., objects which cannot be with certainty classified, by employing the available knowledge, as members of the set or its complement. Obviously rough sets, in contrast to precise sets, Cannot be characterized in terms of information about their elements. With any rough set a pair of precise sets, called the lower and the upper approximation of the rough set, is associated. The lower approximation consists of all objects which surely belong to the set and the upper approximation contains all objects which possibly belong to the set. The difference

between the upper and the lower approximation constitutes the boundary region of the rough set. Approximations are fundamental concepts of rough set theory. Rough set based data analysis starts from a data table called a decision table, columns of which are labeled by *attributes, rows* by objects of interest and entries of the table are at-tribute values. Attributes of the decision table are divided into two disjoint groups called condition and decision attributes, respectively. Each row of a decision table induces a decision rule, which specifies

decision (action, results, outcome, etc.) if some conditions are satisfied. If a decision rule uniquely determines decision in terms of conditions – the decision rule is certain. Otherwise the decision rule is uncertain. Decision rules are closely connected with approximations. Roughly speaking, certain decision rules describe lower approximation of decisions in terms of conditions, whereas uncertain decision rules refer to the boundary region of decisions. With every decision rule two conditional probabilities, called the certainty and the coverage coefficient, are associated.

The certainty coefficient expresses the conditional probability that an object belongs to the decision class specified by the decision rule, given it satisfies conditions of the rule. The coverage coefficient gives the conditional probability of reasons for a given decision. It turns out that the certainty and coverage coefficients satisfy Bayes' theorem. That gives a new look into the interpretation of Bayes' theorem, and offers

PhD scholar, RNTU,
Bhopal

HOW TO CITE THIS ARTICLE:

Harsh Mathur (2018).
Web Data Using
Mining Techniques
with The Help of
Open Source Software
Packages, International
Educational Journal
of Science and
Engineering (IEJSE),
Vol: 1, Issue: 3, 16-19

a new method data to draw conclusions from data. Rough set theory has an overlap with many other theories dealing with imperfect knowledge, e.g., evidence theory, fuzzy sets, Bayesian inference and others. Nevertheless, the theory can be regarded as an independent, complementary – not competing discipline, in its own rights. More information about rough sets and their applications can be found in the references and the Web.

2. ARCHITETURE OF PROPOSED MODEL

Figure 1 shows the proposed model of the system. The model has three-layers - Information Extraction and Web mining, and User Friendly Interface.

The bottom layer is the information extraction and analysis engine which takes a DBLP database [3] which is in XML format, extract information from database and analyse that information.

The middle layer is the functionality means functional module layer, which implements the major functions like profiling, ranking and page accessing. Profiling is the process of getting relevant information about the particular object present in the database. Web accessing is the process of assigning priority to object according to some classifications present in the database.

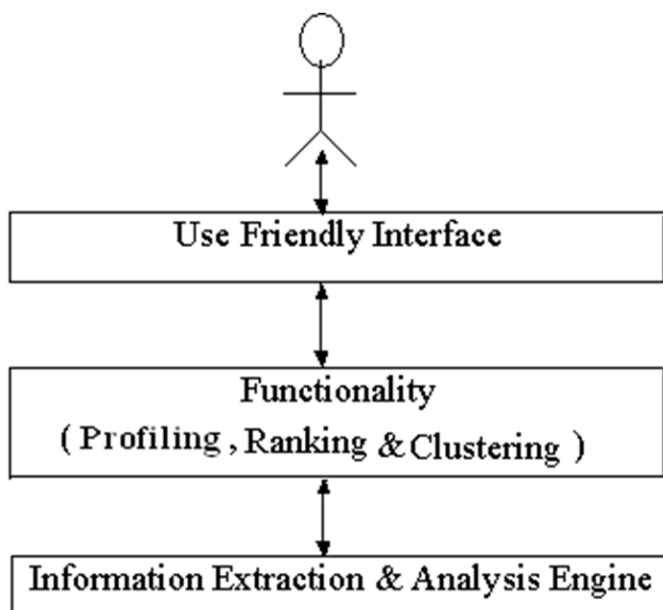


Figure 1: Proposed Model

The top layer is a user-friendly interface, contains a user-friendly and visualization-enhanced interface, which interacts with user and responds to their request from web mining.

3. PROPOSED WORK

A. Previous Work with respect to MBI_A data mining

Different combinations of mining techniques were already suggested for web access recommendation. Pawlak[4] introduced web access prediction model by integrating roughest clustering with Markov model. It has major drawback that lack of prediction accuracy due to approximation while forming clusters. The possibility of an object for belonging to a cluster

can reduce the cluster tightness, which in turn affects prediction accuracy. The sequential mining suggested in that work is all k-th order Markov model. For high order Markov models, if coverage is very less, it affects accuracy of prediction. Hence in this paper it is proposed to do support pruning. By which, the states that have less support or low coverage are eliminated while making recommendations. Another model [5] based on Markov process for web access prediction has drawback of high complexity due to consideration of all access sequences' through out the prediction process. Sometimes, in this approach noisy or irrelevant access sequences participate in web page access prediction affects the accuracy. Hence it is suggested to do pattern discovery before sequential mining. Some works based on association rule were made, where prediction accuracy is affected by contradictory predictions due to the enervation of too many rules and participation of huge number of access sequences in mining process. A combined approach like integrating Markov model and association rule [6] is also affected by mining all types of access sequences i.e) without clustering. In agglomerative clustering using k neighbours, at later stage of clustering process, distant neighbours may also fall into k neighbours and they decrease the cluster tightness. Hence threshold is set to eliminate such objects. Hence, it is proposed to find out highly homogeneous access patterns by pair wise nearest neighbour based clustering. The resultant patterns are highly relevant and the size dataset that are utilized for sequential mining process is highly reduced.

4. DYNAMIC MODELLING APPROACH FOR WEB USAGE MINING AND DEVELOPMENT OF SOFTWARE

Today with such an overwhelming quantity of data available on the internet, the traditional search tools problems appear. Users often suffer from information overload. They have to filter irrelevant information by themselves. The contradiction between rich data and poor knowledge caused the emergence of web data mining. In recent years, due to the rapid development of web, more techniques related to internet and web applications have been introduced in many universities around the world to better equip students to effectively utilize the power of the web. These include basic courses such as Internet networking, Internet application development, Web search engines, and so forth. Web mining defined as the use of data mining, text mining, and information retrieval techniques to extract useful patterns and knowledge fro the web. Most data used for mining is collected from web servers, client's proxy servers, or server data bases, all of which generate noisy data. Because web mining is sensitive to noise, data cleaning methods are necessary. However, building a web mining application or a web services application from scratch is not an easy task that every student could complete in a semester. It is believed that using web APIs can simplify the data acquisition process and enables the students to focus on the application of web mining algorithms. The students, thus focus on the web mining algorithms to build value added applications.

4.1 Data preparation

The web log data is used for mining process is integrated, cleaned and relevant attributes like IP and web page accesses are selected. The click stream is a sequence of mouse click made by

every user. The transactions are generated by eliminating noisy, and very short or very long access sequences.

4.2 Introduction:

Rough set theory provides an approach to approximation of sets that leads to useful forms of granular computing. The underlying concept is to extract to what extent a given set of objects (e.g. extracted feature samples) approximate another set of objects of interest. Rough sets offer an effective approach of managing uncertainties similarity value between every pair of transaction is created. For every transaction, its first k nearest access sequences is identified. Among the whole and can be employed for tasks such as data dependency analysis, feature identification, dimensionality reduction, and pattern classification. Based on rough set theory it is possible to construct a set of simple if-then rules from information tables. Often, these rules can reveal previously undiscovered patterns in sample data. Rough set methods can also be used to classify unknown data based on already gained knowledge. Unlike many other techniques, rough set analysis requires no external parameters and uses only the information present in the input data. Rough set theory can be utilised to determine whether sufficient data for a task is available respectively to extract a minimal sufficient set of features for classification which in turn effectively performs feature space dimensionality reduction. Although, compared to other methods, a relatively recent technique, these characteristics have prompted various rough set approaches in the general domain of medical informatics. In the following we will therefore, after giving a brief introduction to basic rough set concepts, provide an overview of the use of rough sets in this area. In particular, we will show how rough sets have been used for medical image segmentation, classification, for mining medical data, and in medical decision support systems.

4.3 Rough set Theory:

Rough set theory [11, 14] is a fairly recent intelligent technique for managing uncertainty that is used for the discovery of data dependencies, to evaluate the importance

of attributes, to discover patterns in data, to reduce redundancies, and to recognise and classify objects. Moreover, it is being used for the extraction of rules

from databases where one advantage is the creation of readable if-then rules. Such rules have the potential to reveal previously undiscovered patterns in the data; furthermore, it also collectively functions as a classifier for unseen samples. Unlike other computational intelligence techniques, rough set analysis requires no external parameters and uses only the information presented in the given data. One of the useful featur of rough set theory is that it can tell whether the data is complete or not based on the data itself. If the data is incomplete, it will suggest that more information about the objects is required. On the other hand, if the data is complete, rough sets are able to determine whether there are any redundancies and find the minimum data needed for classification. This property of rough sets is very important for applications where domain knowledge is very limited or data collection is expensive because it makes sure

the data collected is just sufficient to build a good classification model without sacrificing accuracy.

Data set: We apply data set

X13	0	0	0	0	0	1
X14	0	0	1	0	0	0
X15	0	0	0	0	0	1
X16	0	0	0	0	0	1
X17	0	0	0	0	0	1
X18	1	1	1	0	0	0
X19	1	1	1	1	0	0
X20	1	0	0	0	1	0
X21	0	1	0	0	1	1
X22	1	1	1	0	0	0
X23	0	0	0	0	0	1
X24	0	1	0	0	1	1
X25	0	1	0	0	1	1
X26	0	0	0	0	0	1
X27	1	1	1	0	0	0
X28	0	1	0	0	0	0
X29	1	1	1	0	0	0
X30	0	1	0	0	1	1
X31	0	1	1	0	1	1
X32	1	0	0	0	1	0
X33	0	1	0	0	1	1
X34	0	1	1	0	1	1
X35	0	1	0	0	0	0

Table 1.1

from Table 1.1. than arrange and apply lower and upper approximation model in this data set. after applying data set we conclude that we get no of more pages from upper approximation model.



Table 1.2: (lower approximation data)

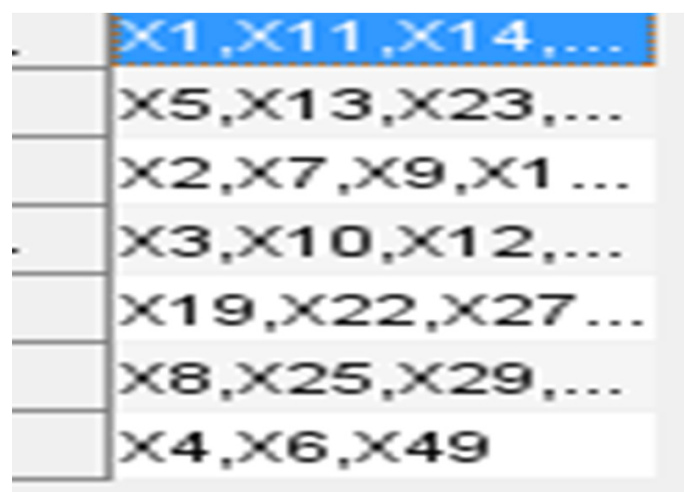


Table 1.3: (upper approximation)

Software Evaluation: We have taken the 200 web pages and

make the similarities among pages by upper approximation. We are implementing this in Mat lab.

Explanation: The discredibility matrix corresponding to the sample database shown in table1 with $U=\{X1,X2,\dots,X7\}$, $C=\{a,b,c,d\}$, $D=\{E\}$ is shown in table 2. $M(X1,x3)=(b,c,d)$

	a	b	c	d	E
X1	1	0	2	1	1
X2	1	0	0	0	1
X3	1	2	2	0	2
X4	2	2	0	1	0
X5	2	1	1	0	2
X6	2	1	2	1	1

Table1.4: A sample database

	X1	X2	X3	X4
X1	-			
X2	bcd	b,c	c,d	
X3	b	b,d	---	a,b,c
X4	a,b,c,d	a,b,c	---	a,b,c
X5	a,b,c	b,c	a,b,c	c,d
X6	---	---	-----

Table 1.5

Reduct are $\{b,c\}$ and $\{b,d\}$ core = $\{b\}$

$PX=\{X1,X2\} \cup \{X4\}$

$PX=\{X1,X2\} \cup \{X4\} \cup \{X3,X7,X10\}$

5. CONCLUSION

The proposed method resulted in good prediction accuracy with less state space complexity. Through these mining projects we observed that mostly real time data through web resources were able to build innovative business applications with in a short period of time. Our proposed method gives up high precision quality as compair to markov chain model.

6. FUTURE WORK

We will explore the software version and on multiple data base

7. REFERENCES

1. Pasi Franti, Olli Virtajoki, and Ville Hautamaki "Fast Agglomerative Clustering Using a k Nearest Neighbor graph", IEEE transaction on pattern analysis and machine intelligence. Vol 28, No 11. November 2006, pp 1875-1880
2. Pasi Franti, Timo Kaukoranta, Day-Fann Shen and Kuo-Shu Chang "Fast and Memory Efficient Implementation of exact PNN", IEEE Transaction on image processing. Vol 9, No 5, May 2000. pp 773-777
3. Mathias G'ery, Hatem Haddad, "Evaluation of Web Usage Mining approaches for user's next request prediction" WIDM '03 Boston, USA, ACM
4. Siripom chimphlee, Naomie Salim, Mohd Salihin Bin Ngadiman, Witcha, Surat, "Rough Sets Clustering and Markov Model for

5. Devanshu Dhyani, Sourav S Bhowmick, Wee-Keong Ng, "Modelling and predicting web page accesses using Markov Processes", IEEE, Computer Society, 2003, 1529-4188
6. Faten Khalil, Jiuyong Li, Hua Wang, "Integrating Recommendation Models for Improved Web Page Prediction Accuracy", Australian Computer Society, 2008, Conferences in Research and Practice in Information Technology, Vol 74.
7. Bamshad Mobasher, Honghua Dai, Tao Luo, Miki Nakagawa, "Effective Personalizaion based on association rule discovery from Web Usage Data", ACM workshop on Web Information and Data management, Nov 2001.
8. Mukund Deshpande and George Karypis, "Selective markov model for predicting web-page accesses", Army High performance Computing Research Center, pp.1-15
9. Faten Khalil, Jiuyong Li, Hua Wang, "Integrating Markov model with clustering for predicting web page accesses", Australian Conference, Mar 2007, pp 1-26
10. Jose Miguel Gago, Carlos Juiz "Web Mining Service (WMS), a public and free Servie for Web Data Mining", IEEE Fourth international Conference on Internet and Web Applications and Services, pp. 351-356, 2009.
11. Li Lan, Rong Qiao-mei "Research of Web Mining Technology Based on XML", IEEE International Conference on Network Security, Wireless Communications and Trusted Computing, Vol.2, pp. 653-656, 2009.
12. Amazon. 2006. Amazon Web Services Website. <http://www.amazon.com/webservice>