

Development of Association Rule Based Prediction Model for Web Documents

¹Sachin Sharma, ²Simple Sharma, ³Anupriya Jain, ⁴Rashmi Aggarwal, ⁵Seema Sharma

^{1,3,4}Dept. of Computer Applications, Manav Rachna International University,
Faridabad, Haryana, India

²Dept. of Engg and Technology, Manav Rachna International University,
Faridabad, Haryana, India

Abstract

The rapid expansion of the WWW has created an unprecedented opportunity to disseminate and gather information online. Electronic Commerce is emerging as the biggest application of WWW. As this trend becomes stronger and stronger, there is much need to study web-user behaviors to better serve the users and increase the value of enterprises. One important data source for this study is the web-log data that traces the user's web browsing actions. From the web logs, one can build prediction models that predict with high accuracy the user's next request based on past behavior. To do this with the traditional association rule methods will cause a number of serious problems due to extremely large data size and the rich domain knowledge that must be applied. Most web log data are sequential in nature and exhibit the "most recent-most important" behavior. To overcome this difficulty, we examine two dimensions of building prediction models. This paper proposes a better overall method for prediction model representation and refinement.

Keywords

Rule-Representation, WWW, Prediction

I. Introduction

The information revolution is generating mountains of data and data everywhere. With rapid growth in size and number of available databases in Commercial, Industrial, Administrative and other applications, it is becoming extremely necessary to examine how to organize the data and extract knowledge from this huge amount of data. A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data Mining, the extraction of hidden information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data Mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. With the explosive growth of information sources available on WWW, it has become increasingly necessary for users to utilize automated tools to find the desired information resources, and to track and analyze their usage patterns. These factors give rise to the necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge from web. Web Mining can be defined as the discovery and analysis of useful information from the WWW. This describes the automatic resources available online. There are three knowledge discovery domains that pertain to web mining: web content Mining, Web Structure Mining and Web Usage Mining. Web Content Mining [7] is an automatic process that goes beyond keyword extraction. Since the content of a text document presents no machine readable semantic, some approached have suggested restructuring the document content in a representation that could be exploited by machines. The usual approach to exploit known structure in documents is to use wrappers to map documents to

some data model. There are two groups of web content mining strategies: those that directly mine the content of documents and those that improve on the content search of other tools like search engines. WWW can reveal more information than just the information contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics. The goal of Web Structure Mining is to generate structural summary about the website and web page. Technically, Web Content Mining mainly focuses on the structure of inner of document, while web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Web Structure mining has a nature relation with the web content mining, since it is very likely that the web documents contain links, and they both use the real or primary data on the web. Web Servers record and accumulate data about user interactions whenever request for resources are received. Analyzing the web access logs of different websites can help understand the user behavior and the web structure, thereby improving the design of this colossal collection of resources. There are two main tendencies in web usage mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking. The general access pattern tracking analyzes the web logs to understand access patterns and trends. These analyses can shed light on better structure and grouping of resource providers. Many web analysis tools existed but they are limited and usually unsatisfactory.

II. Web Mining Techniques

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In some cases, users may have no idea which kinds of patterns in their data may be interesting, and hence may like to search for several different kinds of patterns in parallel. Thus, it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectations or applications. The information gathered through Web Mining is evaluated by using traditional data mining parameters such as:

1. Concept/ Class Description Method
2. Association Analysis Method
3. Classification and Prediction Methods
4. Cluster Analysis

A. Concept/ Class Description Method

Data can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise and yet precise terms. Such descriptions of a class or a concept are called Class/ Concept Descriptions. These descriptions can be derived via Data Characterization and Data Discrimination. Data Characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a database query. Data Discrimination is a comparison of the general features of

target class data objects with the general features of objects from one or a set of contrasting classes. The target and contrasting classes can be specified by the user, and the corresponding data objects retrieved through database queries.

B. Association Rule Method

Association Rule is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association Analysis is widely used for Market basket or transaction data analysis. One of the reasons behind maintaining any database is to enable the user to find the interesting patterns and trends in the data. Association Rule Mining finds interesting association relationships among a large set of data items. With massive amount of data continuously being collected and stored, many industries are becoming interested in mining association rules from their databases. Association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts.

1. Support of an Association Rule

Is defined as the percentage of records that contain XY to the total number of records in database. The count for each item is increased by one every time the item is encountered in different transaction T in database D during the scanning process. It means the support count does not take the quantity of the item into account.

$$\text{Support (XY)} = \frac{\text{Support Count of XY}}{\text{Total no of transaction in D}}$$

Confidence of an Association Rule is defined as the percentage of the number of transactions that contain XY to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule can be generated.

$$\text{Confidence (XY)} = \frac{\text{Support (XY)}}{\text{Support(X)}}$$

Association Rule Mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is decomposed into two sub problems:

To find those item sets whose occurrence exceeds a predefined threshold in the database, those item sets are called Frequent Items sets. To generate association rules from those large item sets with the constraints of minimal confidence.

C. Classification and Prediction Methods

Classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of asset of training data. Classification can be used for predicting the class label of data objects. However, in many applications, users may wish to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data and is often referred to as Prediction. Although Prediction may refer to both data value prediction and class label prediction. Prediction also encompasses the identification of distribution trends based on the available data. For example, a classification model may be build to categorize bank applications as safe or risky, while a prediction model may be build to predict the expenditures of potential customers on computer equipment given their income and occupation.

D. Clustering Method

Unlike classification and prediction, which analyzes class-labeled data objects, clustering analyzes data objects without consulting a known class-label. In general, class labels are not present in training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects with in a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, form which rules can be derived.

III. Prediction Models in Web Mining

With the increasing popularity of WWW, the amount of information, available on the internet has grown exponentially. As a result, the internet has been attracting numerous users as well as heavy network traffic during the last decade. Due to limited bandwidth capacity of the internet, and the increasing demands from a variety of transmission-heavy web based applications, the WWW has been ironically quoted as the 'World Wide Wait'. Furthermore, users and internet content providers are complaining about the lack of customization capacity of server pages for visitors with different browsing habits. This is caused by the web servers. Nowadays, the most popular approaches to solve this problem include recognizing individual users by storing cookies in client browsers, and explicitly asking users to inform the servers about their preferences.

A. Prediction Models

The web-log data consists of a long sequence of URL's requested by different users bearing different IP addresses. These IP addresses can not be used to identify users in many cases. However, they do provide session information where the several consecutive accesses by the same user are grouped in one session. To build a prediction model, a finite window is moved across the web log. At any moment, a current window consists of an antecedent window and a consequent window. The prediction of requests within consequent window is based on the request made within the antecedent window. To apply such a model to web-log data, one can slide the current window through the sequence and map each window to a record. This paper studies the method for building user-independent prediction model from web-log data using association rules. This prediction model, which power the web servers are used to make predictions for any user, not a particular one. We examine important dimensions of building prediction models, namely the antecedents of rules and the criterion for selecting prediction rules.

B. Web Logs and User Sessions

Since our web-document prediction model is based on web-server logs, it is important to understand what information web server logs contain. Web logs are comprised of a collection of chronologically recorded requests from a large number of users to a given website. These web server logs contain million of records, where each record refers to a visit by a user to a certain web page served by a web server. For each visit, web log records remote user's host name or IP addresses, the time when the request arrived, the HTTP method, the remote user used, the URL of the visited web document, the status code of the HTTP response and the number of bytes returned to the user.

Given a web log, the first step is to clean the raw data. Documents that are not requested directly by users are filtered out. These are image or video clips request in the log that are retrieved automatically after accessing request to a document page containing embedded links to these files.

The next step is to extract user sessions from web logs. A user session is a relatively independent sequence of web request from the same user.

C. Rule Representation Methods

We now discuss how to extract rules of the form LHS, RHS from the log table. The RHS in each association rule is the next page requested by the user. However, there is more than one choice to select LHS to extract the information in the antecedent windows. Here, we will adopt a method to extract rule based on different criteria for selecting the LHS.

1. Subsequence Rules

The first rule representation is called the subsequence rule which takes into account the order information in the sessions. A subsequence within the antecedent window is formed by a series of URL's that appear in the same sequential order as they were accessed in the web log data set. However, they do not have to occur right next to each other, nor are they required to end with the antecedent window. When this type of rule is extracted from the log Tables, the left hand side of the rules will include the order information. The following table is an example of extracting subsequence rule from a record in a log Table.

W1	W2	Extracted Rules
A,B,C	D	{A,B,C} D, {A,B} D, {B,C} D, {A,C} D, {A} D, {B} D, {C} D
B,C,D	E	{B,C,D} E, {B,C} E, {C,D} E, {B,D} E, {B} E, {C} E, {D} E

Example: Subsequence Representation Rule Extraction

D. Rule Selection Methods

Our goal is to output only one best guess on a class for a given observation. From the construction of our prediction methods, we conclude that no matter which rule representation method we use, each testing case may give rise to more than one rule whose LHS's match the case. In other words, there may be more than one rule in our models that can apply for the testing case. As a result, we need a way to select among all rules that apply. We call this problem multiple applicable rules.

IV. Proposed Method

A. Outlines

Each Rule-representation method and Rule-Selection method pair gives rise to a prediction model. A question arises which Prediction model is the best in making predictions. Our goal is to select the best Rule Representation method and Rule selection method among all. Given a set of rules, we can make prediction for any web document. To measure the performance of different models, we employ Efficiency as the performance metrics and can be defined as $\text{Efficiency} = C/N$. We set the minimum support threshold value for the experiment as 3% and minimum confidence threshold value as 22.

B. Experimental Setup

Users are identified only by a sequential number, for example, User #10003, User#10009. The file contains no personally identifiable information.

1. Dataset format

The data is in an ASCII-based sparse-data format called "DST" and each line starts with a letter which tells the line's type.

The three line types of interest are:

(i). Attribute lines

For example, 'A, 1277,1,"NetShow for PowerPoint" ,"/stream"' where

'A' marks this as an attribute line

'1277' is the attribute ID number

'NetShow for PowerPoint' is title

"/stream" is the URL.

(ii). Case and Vote Lines

For each user, there is a case line followed by zero or more vote lines. For example:

C,"10164",10164

V,1123,1

V,1009,1

V,1052,1

Following attributes are used in Prediction Model

1. A,1004,1,"Microsoft.com Search","/search"
2. A,1003,1,"Knowledge Base",/kb"
3. A,1008,1,"Free Downloads","/msdownload"
4. A,1001,1,"Support Desktop","/support"
5. A,1009,1,"Windows Family of OSs","/windows"
6. 1,1037,1,"Windows 95","/windows 95"
7. A,1071,1,"Products","/products"
8. A,1034,1,"Internet Explorer","/ie"

C. Training Data

Sample of the data: This is the part of log file showing the Case and Vote Lines.

C,"10001",10001	C,"10016",10016	V,1017,1	V,1038,1
V,1000,1	V,1025,1	V,1004,1	V,1031,1
V,1001,1	V,1026,1	C,"10023",10023	V,1052,1
V,1002,1	C,"10017",10017	V,1008,1	V,1053,1
C,"10002",10002	V,1027,1	C,"10024",10024	V,1018,1
V,1001,1	V,1017,1	V,1044,1	C,"10036",10036
V,1003,1	V,1026,1	C,"10025",10025	V,1051,1
C,"10003",10003	V,1028,1	V,1045,1	V,1054,1
V,1001,1	C,"10018",10018	C,"10026",10026	V,1018,1
V,1003,1	V,1004,1	V,1034,1	V,1035,1
V,1004,1	C,"10019",10019	C,"10027",10027	V,1008,1
C,"10004",10004	V,1071,1	V,1008,1	V,1009,1
V,1005,1	V,1004,1	V,1046,1	V,1026,1
C,"10005",10005	V,1018,1	V,1034,1	V,1040,1
V,1006,1	V,1029,1	C,"10028",10028	V,1052,1
C,"10006",10006	V,1008,1	V,1295,1	V,1041,1
V,1003,1	V,1030,1	C,"10029",10029	V,1003,1
V,1004,1	V,1031,1	V,1034,1	V,1034,1
C,"10007",10007	V,1032,1	C,"10030",10030	V,1048,1
V,1007,1	V,1003,1	V,1017,1	C,"10037",10037
C,"10008",10008	V,1033,1	V,1048,1	V,1008,1
V,1004,1	V,1002,1	C,"10031",10031	V,1055,1

C,"10009",10009	C,"10020",10020	V,1045,1	V,1056,1
V,1008,1	V,1008,1	V,1049,1	V,1017,1
V,1009,1	V,1001,1	V,1018,1	V,1032,1
C,"10010",10010	V,1034,1	V,1008,1	C,"10038",10038
V,1010,1	V,1002,1	V,1035,1	V,1008,1
V,1000,1	C,"10021",10021	V,1027,1	V,1027,1
V,1011,1	V,1017,1	V,1046,1	V,1026,1
V,1012,1	V,1004,1	V,1009,1	V,1041,1
V,1013,1	V,1018,1	V,1031,1	V,1032,1
V,1014,1	V,1035,1	V,1041,1	V,1001,1
C,"10011",10011	V,1036,1	V,1001,1	V,1003,1
V,1015,1	V,1008,1	V,1003,1	V,1018,1
V,1016,1	V,1037,1	V,1002,1	V,1057,1
V,1017,1	V,1009,1	V,1034,1	C,"10039",10039
V,1018,1	V,1038,1	C,"10032",10032	V,1000,1
V,1019,1	V,1026,1	V,1050,1	V,1058,1
C,"10012",10012	V,1039,1	C,"10033",10033	V,1017,1
V,1020,1	V,1040,1	V,1032,1	V,1049,1
V,1021,1	V,1032,1	C,"10034",10034	V,1001,1
C,"10013",10013	V,1041,1	V,1037,1	V,1034,1
V,1022,1	V,1042,1	V,1009,1	C,"10040",10040
C,"10014",10014	V,1034,1	V,1004,1	V,1008,1
V,1023,1	V,1043,1	C,"10035",10035	V,1034,1
C,"10015",10015	C,"10022",10022	V,1008,1	
V,1024,1	V,1008,1	V,1051,1	

D. Testing Data –Log Files

Sample of testing data: Two samples of testing are shown.

C,"10001",10001	V,1031,1	V,1018,1	C,"10048",10048
V,1038,1	V,1003,1	C,"10019",10019	V,1008,1
V,1026,1	V,1034,1	V,1034,1	V,1010,1
V,1034,1	V,1018,1	C,"10020",10020	C,"10049",10049
C,"10002",10002	C,"10008",10008	V,1008,1	V,1017,1
V,1008,1	V,1065,1	V,1004,1	C,"10050",10050
V,1056,1	V,1123,1	V,1034,1	V,1000,1
V,1032,1	V,1009,1	C,"10021",10021	V,1004,1
C,"10003",10003	V,1007,1	V,1207,1	C,"10051",10051
V,1064,1	C,"10009",10009	C,"10022",10022	V,1119,1
V,1065,1	V,1017,1	V,1038,1	V,1009,1
V,1020,1	V,1043,1	V,1125,1	C,"10052",10052
V,1007,1	C,"10010",10010	V,1026,1	V,1008,1
V,1038,1	V,1032,1	V,1053,1	C,"10053",10053
V,1026,1	V,1004,1	C,"10023",10023	V,1020,1
V,1052,1	C,"10011",10011	V,1008,1	V,1004,1
V,1041,1	V,1036,1	V,1073,1	V,1195,1
V,1028,1	V,1077,1	C,"10024",10024	C,"10054",10054
C,"10004",10004	V,1003,1	V,1073,1	V,1053,1
V,1004,1	V,1001,1	C,"10025",10025	C,"10055",10055
C,"10005",10005	C,"10012",10012	V,1004,1	V,1008,1
V,1017,1	V,1004,1	C,"10026",10026	V,1034,1
V,1156,1	C,"10013",10013	V,1009,1	C,"10056",10056
V,1004,1	V,1008,1	C,"10027",10027	V,1004,1
V,1018,1	V,1130,1	V,1008,1	C,"10057",10057
V,1008,1	V,1035,1	V,1073,1	V,1035,1
V,1027,1	V,1034,1	C,"10043",10043	V,1001,1
V,1009,1	V,1018,1	V,1008,1	V,1003,1
V,1046,1	C,"10014",10014	V,1034,1	V,1002,1
V,1038,1	V,1008,1	V,1004,1	V,1004,1
V,1006,1	V,1009,1	C,"10044",10044	V,1018,1
V,1026,1	V,1001,1	V,1001,1	C,"10058",10058

V,1052,1	V,1034,1	C,"10045",10045	V,1008,1
V,1060,1	C,"10015",10015	V,1008,1	V,1000,1
V,1001,1	V,1003,1	C,"10046",10046	V,1046,1
V,1041,1	V,1004,1	V,1030,1	V,1054,1
V,1034,1	C,"10016",10016	V,1037,1	V,1049,1
C,"10006",10006	V,1025,1	V,1009,1	V,1032,1
V,1034,1	V,1026,1	C,"10047",10047	V,1001,1
V,1004,1	C,"10017",10017	V,1008,1	V,1003,1
C,"10007",10007	V,1004,1	V,1035,1	V,1034,1
V,1008,1	C,"10018",10018	V,1037,1	V,1018,1
V,1000,1	V,1008,1	V,1009,1	C,"10059",10059
V,1035,1	V,1058,1	V,1070,1	V,1008,1
V,1016,1	V,1017,1	V,1018,1	V,1017,1

E. Development of Prediction Model

The steps involved in this process are:

1. let the attributes be denoted by alphabets

1004 - A 1009 - F
 1003 - B 1037 - G
 1008 - C 1017 - H
 1001 - E 1034 - K

2. Generate the sequence of occurrences of attributes in Log data

AC	BKA	GFK	GFHA
AK	CAK	GHE	GFHE
BA	CBA	HAK	GFHK
BE	CBE	HBA	HACG
BK	CDE	HEB	HBAK
CA	CEA	HEK	HBEK
CB	CEB	HKA	HBKA
CE	CEK	BHEK	HEBA
CF	CFA	CBAK	HEBK
CG	CFB	CBEA	CFEBK
CH	CFE	CBKA	CFHEA
CK	CFH	CFEA	CFHEB
EA	CFK	CFHA	CGFAK
EB	CGF	CFHE	CGFEA
EC	CGH	CFHK	CGFHE
EK	CHA	CGFB	CGFHK
FA	CHK	CGFE	CGFKA
FB	CKA	CGFH	CHEKA
FE	EBA	CGFK	GFHBA
FH	EBK	CHAK	GFHEA
FK	FBA	CHEB	GFHBE
GF	FBE	CHKA	HACFE
HA	FBK	EBKA	HACGF
HB	FEA	EFBEK	HAGFE
HC	FEB	FEAK	HEBKA
HE	FHA	FEBA	CGFHEA
HK	FHE	GFBE	GFHBEA
KA	FHK	GFBK	GFHEBA
BAK	GFA	GFEA	
BEA	GFB	GFEB	
BEK	GFH	GFEK	

3. Generate the windows corresponding to sequences

W1	W2		W1	W2
CBE	A		CGF	E
CBK	A		FHB	E
CFE	A		GFB	E
CFH	A		GFH	E
CGF	A		HAC	G
CHK	A		CGF	H
EBK	A		BHE	K
FEB	A		CBA	K
FHB	A		CFH	K
FHE	A		CGF	K
GFE	A		CHA	K
GFH	A		CHE	K
GFK	A		FEA	K
HBK	A		FEB	K
HEB	A		GFA	K
HEK	A		GFB	K
CFE	B		GFE	K
CGF	B		GFH	K
CHE	B		HBA	K
FHE	B		HBE	K
GFE	B		HEB	K
GFH	B			
CFH	E			

4 (a): For subsequence Rule Representation method, generate the rules for attributes in window

Rules		Rules		Rules
CBE-A		CFH-A		EBK-A
CBK-A		CGF-A		FEB-A
CFE-A		CHK-A		FHB-A

Rules		Rules		Rules		Rules
FHE-A		HEB-K		FB-E		HE-K
GFE-A		BE-A		FH-E		B-A
GFH-A		BK-A		GB-E		C-A
GFK-A		CB-A		GF-E		E-A
HBK-A		CE-A		GH-E		F-A
HEB-A		CF-A		HB-E		G-A
HEK-A		CG-A		AC-G		H-A
CFE-B		CH-A		HA-G		K-A
CGF-B		CK-A		HC-G		C-B
CHE-B		EB-A		CF-H		E-B
FHE-B		RK-A		CG-H		F-B
GFE-B		FB-A		GF-H		G-B
GFH-B		EFE-A		BA-K		H-B
CFH-E		FH-A		BE-K		B-E
CGF-E		FK-A		BH-K		C-E
FHB-E		GE-A		CA-K		F-E
GFB-E		GF-A		CB-K		G-E
GFH-E		GK-A		CE-K		H-E
HAC-G		HB-A		CF-K		A-G

CGF-H		HE-A		CG-K		C-G
BHE-K		HK-A		CH-K		H-G
CBA-K		CE-B		EA-K		C-H
CFH-K		CF-B		EB-K		F-H
CGF-K		CG-B		FA-K		G-H
CHA-K		CH-B		FB-K		A-K
CHE-K		FE-B		FE-K		B-K
FEA-K		FH-B		FH-K		C-K
FEB-K		GE-B		GA-K		E-K
GFA-K		GF-B		GB-K		F-K
GFB-K		GH-B		GE-K		G-K
GFE-K		HE-B		GF-K		H-K
GFH-K		CF-E		GH-K		
HBA-K		CG-E		HA-K		
HBE-K		CH-E		HB-K		

5 (a): Arrange the rule alphabetically and calculate the support for each

Rules	Support	Rules	Support	Rules	Support
CBE-A	0.7	CFH-A	0.7	EBK-A	0.7
CBK-A	0.7	CGF-A	0.7	FEB-A	0.7
CFE-A	0.7	CHK-A	0.7	FHB-A	0.7

Rules	Support	Rules	Support	Rules	Support	Rules	Support
FHE-A	0.7	HEB-K	0.7	FB-E	1.41	HE-K	2.82
GFE-A	0.7	BE-A	1.41	FH-E	3.52	B-A	7.75
GFH-A	0.7	BK-A	2.82	GB-E	0.7	C-A	1.41
GFK-A	0.7	CB-A	1.41	GF-E	3.52	E-A	8.45
HBK-A	0.7	CE-A	0.7	GH-E	0.7	F-A	4.23
HEB-A	0.7	CF-A	0.7	HB-E	2.11	G-A	2.11
HEK-A	0.7	CG-A	0.7	AC-G	1.41	H-A	6.34
CFE-B	0.7	CH-A	1.41	HA-G	0.7	K-A	9.15
CGF-B	0.7	CK-A	0.7	HC-G	0.7	C-B	4.23
CHE-B	0.7	EB-A	2.11	CF-H	2.82	E-B	11.27
FHE-B	0.7	EK-A	1.41	CG-H	0.7	F-B	6.34
GFE-B	0.7	FB-A	0.7	GF-H	4.23	G-B	2.82
GFH-B	0.7	FE-A	2.82	BA-K	2.11	H-B	9.15
CFH-E	0.7	FH-A	2.11	BE-K	1.41	B-E	7.04
CGF-E	0.7	FK-A	1.41	BH-K	0.7	C-E	2.82
FHB-E	0.7	GE-A	0.7	CA-K	0.7	F-E	10.56
GFB-E	0.7	GF-A	2.11	CB-K	0.7	G-E	3.52
GFH-E	0.7	GK-A	0.7	CE-K	0.7	H-E	11.97
HAC-G	0.7	HB-A	2.11	CF-K	0.7	A-G	1.41
CGF-H	0.7	HE-A	1.41	CG-K	0.7	C-G	9.86
BHE-K	0.7	HK-A	1.41	CH-K	1.41	H-G	0.7
CBA-K	0.7	CE-B	0.7	EA-K	1.41	C-H	6.34
CFH-K	0.7	CF-B	0.7	EB-K	2.82	F-H	13.38
CGF-K	0.7	CG-B	0.7	FA-K	1.41	G-H	3.52
CHA-K	0.7	CH-B	0.7	FB-K	1.41	A-K	8.45
CHE-K	0.7	FE-B	3.52	FE-K	1.41	B-K	9.15
FEA-K	0.7	FH-B	2.82	FH-K	2.11	C-K	1.41
FEB-K	0.7	GE-B	0.7	GA-K	0.7	E-K	8.45
GFA-K	0.7	GF-B	2.82	GB-K	0.7	F-K	4.23
GFB-K	0.7	GH-B	0.7	GE-K	0.7	G-K	2.11
GFE-K	0.7	HE-B	3.52	GF-K	2.11	H-K	6.34
GFH-K	0.7	CF-E	2.11	GH-K	0.7		
HBA-K	0.7	CG-E	0.7	HA-K	1.41		
HBE-K	0.7	CH-E	2.11	HB-K	1.41		

6 (a): Eliminating the rules that do not satisfy the minimum support threshold value. Only the rules with support more than 3.2 are bold

Rules	Support		Rules	Support
FE-B	3.52		B-E	7.04
HE-B	3.52		F-E	10.56
FH-E	3.52		G-E	3.52
GF-E	3.52		H-E	11.97
GF-H	4.23		C-G	9.86
B-A	7.75		C-H	6.34
E-A	8.45		F-H	13.38
F-A	4.23		G-H	3.52
H-A	6.34		A-K	8.45
K-A	9.15		B-K	9.15
C-B	4.23		E-K	8.45
E-B	11.27		F-K	4.23
F-B	6.34		H-K	6.34
H-B	9.15			

7 (a): Choose the rule with the highest support among all LHS Di rules. Only the rules with highest support are bold

Rules	Support
FE-B	3.52
HE-B	3.52
FH-E	3.52
GF-H	4.23
K-A	9.15
E-B	11.27
G-E	3.52
H-E	11.97
C-G	9.86
F-H	13.38
G-H	3.52
A-K	8.45
B-K	9.15

8 (a): Calculate the Confidence

Rules	Confidence	Support
FE-B	33.33	3.52
HE-B	29.41	3.52
FH-E	26.32	3.52
GF-H	28.57	4.23
K-A	22.41	9.15
E-B	27.12	11.27
G-E	10.42	3.52
H-E	29.82	11.97
C-G	93.33	9.86
F-H	34.55	13.38
G-H	10.42	3.52
A-K	20.34	8.45
B-K	22.81	9.15

9 (a): Eliminating the rules that do not satisfy the minimum confidence threshold value (<22%)

Rules	Confidence	Support
FE-B	33.33	23.47
HE-B	29.41	20.71
FH-E	26.32	18.53
GF-H	28.57	20.12
K-A	22.41	15.78
E-B	27.12	19.10
H-E	29.82	21
C-G	93.33	65.73
F-H	34.55	24.33
B-K	22.81	16.06

10 (a): Eliminating Redundant and Inconsistent Rules

Rules		Rules
K-A		C-G
E-B		F-H
H-E		B-K

4 (b): For latest-subsequence Rule representation method, generate the rules for the attributes in the window

Rules	Rules	Rules	Rules
CBE-A	CBA-K	HB-A	FA-K
CBK-A	CFH-K	HE-A	FB-K
CFE-A	CGF-K	HK-A	FE-K
CFH-A	CHA-K	CE-B	FH-K
CGF-A	CHE-K	CF-B	GA-K
CHK-A	FEA-K	FE-B	GB-K
EBK-A	FEB-K	FH-B	GE-K
FEB-A	GFA-K	GE-B	GF-K
FHB-A	GFB-K	GF-B	GH-K
FHE-A	GFE-K	GH-B	HA-K
GFE-A	GFH-K	HE-B	HB-K
GFH-A	HBA-K	CF-E	HE-K
GFK-A	HBE-K	CH-E	B-A
HBK-A	HEB-K	FB-E	E-A
HEB-A	BE-A	FH-E	F-A
HEK-A	BK-A	GB-E	H-A
CFE-B	CE-A	GF-E	K-A
CGF-B	CF-A	GH-E	E-B
CHE-B	CH-A	HB-E	F-B
FHE-B	CK-A	AC-G	H-B
GFE-B	EB-A	HC-G	B-E
GFH-B	EK-A	CF-H	F-E
CFH-E	FB-A	GF-H	H-E
CGF-E	FE-A	BA-H	C-G
FHB-E	FH-A	BE-K	F-H
GFB-E	FK-A	CA-K	A-K
GFH-E	GE-A	CE-K	B-K
HAC-G	GF-A	CH-K	E-K
CGF-H	GH-A	EA-K	F-K
BHE-K	GK-A	EB-K	H-K

5 (b): Arrange the rules alphabetically and calculate the support

Rules	Support	Rules	Support	Rules	Support	Rules	Support
CBE-A	0.83	CBA-K	0.83	HB-A	2.5	FA-K	1.67
CBK-A	0.83	CFH-K	0.83	HE-A	1.67	FB-K	1.67
CFE-A	0.83	CGF-K	0.83	HK-A	1.67	FE-K	1.67
CFH-A	0.83	CHA-K	0.83	CE-B	0.83	FH-K	2.5
CGF-A	0.83	CHE-K	0.83	CF-B	0.83	GA-K	0.83
CHK-A	0.83	FEA-K	0.83	FE-B	4.17	GB-K	0.83
EBK-A	0.83	FEB-K	0.83	FH-B	3.33	GE-K	0.83
FEB-A	0.83	GFA-K	0.83	GE-B	0.83	GF-K	2.5
FHB-A	0.83	GFB-K	0.83	GF-B	3.33	GH-K	0.83
FHE-A	0.83	GFE-K	0.83	GH-B	0.83	HA-K	1.67
GFE-A	0.83	GFH-K	0.83	HE-B	4.17	HB-K	1.67
GFH-A	0.83	HBA-K	0.83	CF-E	2.5	HE-K	3.33
GFK-A	0.83	HBE-K	0.83	CH-E	2.5	B-A	8.33
HBK-A	0.83	HEB-K	0.83	FB-E	1.67	E-A	10
HEB-A	0.83	BE-A	1.67	FH-E	4.17	F-A	5
HEK-A	0.83	BK-A	3.33	GB-E	0.83	H-A	7.5
CFE-B	0.83	CE-A	0.83	GF-E	4.17	K-A	10.83
CGF-B	0.83	CF-A	0.83	GH-E	0.83	E-B	13.33
CHE-B	0.83	CH-A	1.67	HB-E	2.5	F-B	7.5
FHE-B	0.83	CK-A	0.83	AC-G	1.67	H-B	10
GFE-B	0.83	EB-A	2.5	HC-G	0.83	B-E	8.33
GFH-B	0.83	EK-A	1.67	CF-H	3.33	F-E	12.5
CFH-E	0.83	FB-A	0.83	GF-H	5	H-E	14.17
CGF-E	0.83	FE-A	3.33	BA-H	2.5	C-G	2.5
FHB-E	0.83	FH-A	2.5	BE-K	1.67	F-H	15.83
GFB-E	0.83	FK-A	1.67	CA-K	0.83	A-K	10
GFH-E	0.83	GE-A	0.83	CE-K	0.83	B-K	10
HAC-G	0.83	GF-A	2.5	CH-K	1.67	E-K	10
CGF-H	0.83	GH-A	0.83	EA-K	1.67	F-K	4.17
BHE-K	0.83	GK-A	0.83	EB-K	3.33	H-K	6.67

6 (b): Eliminate the rules that do not satisfy the minimum support threshold value (i.e. <3.2)

Rules	Support	Rules	Support
BK-A	3.33	H-A	7.5
FE-A	3.33	K-A	10.83
FE-B	4.17	E-B	13.33
FH-B	3.33	F-B	7.5
GF-B	3.33	H-B	10
HE-B	4.17	B-E	8.33
FH-E	4.17	F-E	12.5
GF-E	4.17	H-E	14.17
CF-H	3.33	F-H	15.83
GF-H	5	A-K	10
EB-K	3.33	B-K	10
HE-K	3.33	E-K	10
CB-A	8.33	F-K	4.17
E-A	10	H-K	6.67
F-A	5		

7 (b): Choose the rule with the highest support among all LHS Di rules. Only the rules with highest support are bold

Rules	Support
BK-A	3.33
FE-B	4.17
HE-B	4.17
FH-E	4.17
CF-H	3.33
GF-H	5
EB-K	3.33
K-A	10.83
E-B	13.33
H-E	14.17
F-H	15.83
A-K	10
B-K	10

8 (b): Calculate the Confidence

Rules	Confidence	Support
BK-A	33.33	3.33
FE-B	33.33	4.17
HE-B	29.41	4.17
FH-E	26.32	4.17
CF-H	44.44	3.33
GF-H	28.57	5
EB-K	25	3.33
K-A	25	10.83
E-B	28.57	13.33
H-E	33.33	14.17
F-H	35.19	15.83
A-K	22.22	10
B-K	24	10

9 (b): Eliminating the rules that do not satisfy the minimum confidence threshold value (i.e. <22%)

Rules	Confidence	Support
BK-A	33.33	3.33
FE-B	33.33	4.17
HE-B	29.41	4.17
FH-E	26.32	4.17
CF-H	44.44	3.33
GF-H	28.57	5
EB-K	25	3.33
K-A	25	10.83
E-B	28.57	13.33
H-E	33.33	14.17
F-H	35.19	15.83
A-K	22.22	10
B-K	24	10

10 (b): Eliminating Redundant and inconsistent rules

Rules		Rules		Rules
K-A		H-E		A-K
E-B		F-H		B-K

V. Results of Prediction Model

Following are the results of applying the Rule Representation method on the training data. The value of Average Efficiency of all the rules generated by method by testing on test data1 is shown in the table:

Table 1: Efficiency of Subsequence Rule Representation method

Rules	Testing on Test data 1		
	Y	N	Efficiency
K A	30	0	100
E B	51	44	53.68
H E	19	49	27.94
C G	13	184	6.60
F H	29	39	42.65
B K	10	58	14.71
Average			40.93

Table 2: Efficiency of Latest Subsequence Rule Representation method

Rules	Testing on Test data 1		
	Y	N	Efficiency
K A	30	0	100
E B	51	44	53.68
H E	19	49	27.94
C G	29	39	42.65
F H	33	12	73.33
B K	10	58	14.71
Average			52.05

The following fig. shows the efficiency in %.

M1: Subsequence Rule Representation Method (40.93%)

M2: Latest-Subsequence Rule Representation Method (52.05%)

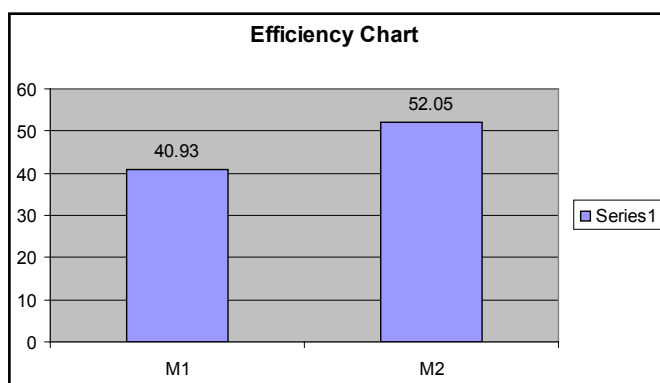


Fig. 1: Efficiency Graph for testing data1

The analysis shows that M2 method is having maximum efficiency of having 0.60%

VI. Conclusion and future Enhancement Scope

In this paper, we have proposed and analyzed Association rules for web document prediction model construction. Our analysis is based on a spectrum of rule-representation methods including subsequence association rule, latest subsequence association rules. In this, We studied the relative performance of different prediction models using real web logs as training and testing data. From comparison of different rules, it is to be concluded that the Latest Subsequence rule representation method uses the most domain knowledge and is suitable for web document prediction.

References

- [1] Jiawei Han, Micheline Kamber, "Data mining, concepts and techniques".
- [2] Danuta Zakrzewska, Ernestina Menasalvas, Liliana Byczkowska-Lipinska, "Methods and Supporting Technologies for Data Analysis".
- [3] [Online] Available: http://www.indianmba.com/Faculty_Column/FC69/fc69.html Association Rule- A Tool For Data Mining
- [4] Tianyi Li, Ke Wang, "Building Association-Rule Based Sequential Classifiers for Web-document ".
- [5] Tianyi Li, "Web-Document Prediction And Presending Using Association Rule Sequential Classifiers".
- [6] Srikant R, Agarwal R, "Mining Sequential Patterns, Generalizations and Performance Improvements", Research Report.
- [7] [Online] Available: <http://www.galeas.de/webmining.html>.
- [8] Morgan Kaufmann, "Mining The Web", Discovering knowledge from hypertext data by: soumen Chakrabarti Publisher.



Sachin Sharma is currently working as Asstt Professor in Computer Science, Manav Rachna International University, Faridabad, India. He is working towards his Ph.D in Operations Research at MDU, Rohtak, India. His qualifications are M.Tech(IT), M.Phil (Computer), MCA and M.Sc(Operations Research). He has 13+ years experience of teaching undergraduate and post graduate students.

He has published 6 research papers in International conferences and participated in many National/ International Conferences. His research interests are in the areas of Software Engg, Data Structures and Operations Research. He has also written various books on "Basics of Data Structure" for BCA, MCA students, "Computer Architecture" for MCA and "Programming in C" for GJU, Hisar. He is a member of CSI and a reviewer member of IJEST, TOJET.



Simple Sharma is currently working as Asstt Professor in Computer Science department, Faculty of Engineering, Manav Rachna International University, Faridabad, India. Her qualifications are M.Tech(CSE). She has 9+ years experience of teaching undergraduate and post graduate students.

She has published 1 research paper in International Journal and participated in many National/ International Conferences. Her research interests are in the areas of Computer Graphics, Operating Systems and System Programming and Administration.



Anupriya Jain is currently working as Asstt Professor in Computer Science, Manav Rachna International University, Faridabad, India. Her qualifications are M.Tech (IT), M.Phil (Computer) and DOEACC B Level. She has 9+ years experience of teaching undergraduate and post graduate students. She has published 6 research papers in International conferences and participated in many National/ International Conferences to

her credit. Her research interests are in the areas of Software Engg and Computer Organization. She has also written two books on "Basics of Data Structure" for BCA, MCA students and "Computer Architecture" for MCA.



Rashmi Aggrawal is currently working as Asstt Professor in Computer Science, Manav Rachna International University, Faridabad, India. Her qualifications are M.Tech (IT), M.Phil (Computer) and DOEACC B Level. She has 9+ years experience of teaching undergraduate and post graduate students. She has published 3 research papers in International conferences and participated in many National/ International Conferences to

her credit. Her research interests are in the areas of Operating System and Artificial Intelligence.



Seema K. Sharma is currently working as Asstt Professor in Computer Science, Manav Rachna International University, Faridabad, India. Her qualifications are M.Phil(Computer) and MCA. She has 5+ years experience of teaching undergraduate and post graduate students. She has published 6 research papers in International conferences and participated in many National/ International Conferences. Her research

interests are in the areas of Software Engg, Data Structures and Java. She has also written books on "Basics of Data Structure" for BCA, MCA students, "Computer Architecture" and Computer Fundamentals & IT Tools" for MCA.