Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 49 (2012) 9852-9857

ISSN: 2229-712X

Network proxy log mining: association rule based security and performance enhancement for proxy server

Sayed Md.Sakib Hossain, S.M. Monzurur Rahman and Md. Faisal Kabir

Department of Computer Science and Engineering, United International University House 80, Road No 8/A, Dhadi. Dhaka 1209, Bangladesh.

ARTICLE INFO

Article history:

Received: 13 June 2012;
Received in revised form:
23 July 2012;
Accepted: 3 August 2012;

Keywords

Log mining;
Data Mining;
DM; Proxy server Log;
Association Rule;
AR; Rule mining;
Network Security.

ABSTRACT

Network Proxy Logs contain useful user access patterns that are waiting to be discovered. By analyzing those logs, it is possible to discover various kinds of knowledge, which can then be applied to improve the performance of proxy server. Association Rule mining, by using Proxy logs, aims to discover interesting user access patterns. This paper proposes a novel approach for proxy log mining. In our approach, the Apriori Algorithm is used to extract important or useful Rules from proxy server ACCESS logs. Our paper's aim is to mine patterns from the Network Proxy Logs and show the difference that some unauthorized clients somehow getting access to information and some authorized clients are not getting access to information. Clients who are unauthorized might be an intruder.

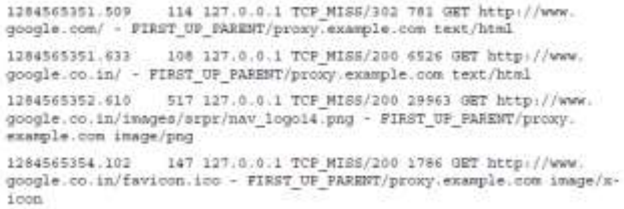
Introduction

Nowadays networking is the simplest and easiest way of information-sharing mechanism in an organization. Everyone is looking for their desired information through the web with a specific threshold set in the proxy server, meaning that the client can only access that type of information which is allowed by the network Administrator. The Administrator can set a proxy server in between the client side and the web server. Administrator may block certain sites like Social networking website such as Facebook.com or twitter.com, Video streaming website like Youtube.com, online shopping site like amazon.com or specific mailing site like gmail.com or yahooomial.com etc.

The explosive growth of the Web has imposed a heavy demand on networking resources and Web servers. Users often experience long and unpredictable delays when retrieving Web pages from remote sites [1]. Hence, an undoubted solution to improve the excellence of Web services would be the increase of network bandwidth, but this may involve increasing cost for the organization. So it is better to look out the criteria of client who are trying to access illegally those types of sites which are already blocked by the network admin. A DM (Data Mining) Technique such as AR (Association Rule) would be feasible to extract the client criteria, who are trying to breach the network [2].

DM sometimes said to be knowledge discovery in databases which is a process of extracting of hidden, previously not known and probable useful information from a large volume of data sets[3][4][5]. The gained information often referred as knowledge of the formed rules, constraints and regularities. AR mining is one of crucial tasks in DM where the rules provide brief statement of possibly important information that can be easily understood by the end users [5] .Researchers have been using many techniques for rule mining such as statistical, AI, decision tree, database, cognitive, etc.

The early work in Proxy Log mining is about searching for some information in browser cache and searching each and every proxy log data in the server log to generate rules and trying to gain Some Information about user criteria. The process is lengthy and cannot be deepen insight to extract ever wondering valuable Information. To Achieve a Better Server performance as well as ensure the security of the network a better approach is proposed which is are: At first, searching For ARs in the Data set and Look For Patterns of Clients who are Getting access to proxy server where They are not supposed to get Access. In the meantime look for Patterns of Clients who are not getting access to proxy server where they are supposed to get Access. Then apply the rules that are getting Access to who are not getting Access and apply the rules that are not getting Access to who are getting Access.



```
1284565351.509    114 127.0.0.1 TCP_MISS/302 781 GET http://www.
google.com/ - FIRST_UP_PARENT/proxy.example.com text/html

1284565351.633    108 127.0.0.1 TCP_MISS/200 6526 GET http://www.
google.co.in/ - FIRST_UP_PARENT/proxy.example.com text/html

1284565352.610    517 127.0.0.1 TCP_MISS/200 29963 GET http://www.
google.co.in/images/srpr/nav_logo14.png - FIRST_UP_PARENT/proxy.
example.com image/png

1284565354.102    147 127.0.0.1 TCP_MISS/200 1786 GET http://www.
google.co.in/favicon.ico - FIRST_UP_PARENT/proxy.example.com image/x-
icon
```

Fig 1.1: ACCESS Log Sample From squid proxy server

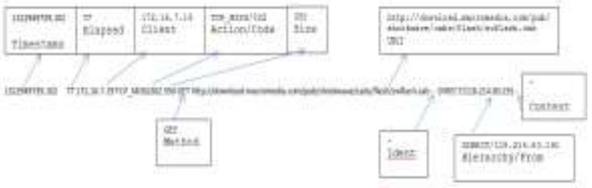


Fig 1.2: Access Log Format

Tele:
E-mail addresses: sakib.diit@gmail.com

This paper is organized as follows. The existing techniques to extracting AR mining with proxy log data are described in Section 2 as a background study. In Section 3, the problem statement is described with real life example. Section 4 describes the problem solution and algorithms to discover ARs. In Section 5 Experiments are conducted to test the proposed rule mining algorithm and results are also reported in. Finally, the summary of this paper is provided in Section 6.

Background study

Proxy server can be used in two ways - Prefetching and caching which improves the performance and security Enhancement of the Web access and is an important component of the Web infrastructure. Nowadays, a number of commercial systems implement some form of prefetching and cashing. For example, a number of browser extensions for Firefox, Netscape and Microsoft Internet Explorer as well as some personal proxies carry out prefetching and cashing [6].In this section, we further present the enthusiasm and involvement of this Related type of work. The earliest work in proxy server log mining is clustering-based prefetching scheme on a Web Cache environment describing Web prefetching and cashing that utilizes the spatial locality of Web objects which will cause significant improvements on the performance of the Web infrastructure by using data mining techniques[7].

Baoyao Zhou, Siu Cheung Hui and Kuiyu Chang [8] conclude that by using Formal Concept Analysis (FCA) data analysis method, proxy log usage mining aims to discover interesting user access patterns from network proxy logs. Applying this FCA they mine association rules that are constructed from web proxy logs. On the other hand FCA rules are also applied to compare the performance with that of classical Apriori-mined rules. The experimental results shows that the FCA approach not only generates far fewer rules than Apriori-based algorithms, the generated FCA rules are also of comparable with respect to three objective performance measures.

DM technique, to mine network traffic log, based on its frequency and Filtering-Rule Generalization (FRG), not only reducing the number of policy rules but also identify any decaying rule and a set of few dominant rules. And it also generates a new set of efficient log policy rules. As a result of these mechanisms, network security administrators can automatically review and update the rules [9]. Some Work related to Anomaly extraction meta-data are also carried out to identify suspicious flows and apply association rule mining to those meta-data to find and summarize the event flows. By using rich traffic data from a backbone network can reduce the classification cost, in terms of items (flows or rules) that need to be classified. The technique effectively isolate event flows in all analyzed cases [10] .Another method to solve a problem of caching non-homogenous Web objects is also carried out that differs from traditional caching. This shows a new algorithm for caching policies designed for Web objects that can be regarded as a generalization of the standard LRU (Least Recently Used) algorithm and it can examine the performance of Web caching algorithms via event- and trace-driven simulation [11].

Though earlier work has used simplistic actions for determining method for the transaction boundaries, and has not addressed the problem of interleaving and noisy transactions with a simplistic view that can lead to poor performance in building models to predict future access patterns. Thus Wenwu Lou, Guimei Liu, Hongjun Lu, Qiang Yang present a more advanced cut-and-pick method for determining the access transactions from proxy logs which can decide on more reasonable transaction boundaries and can remove noisy accesses.it also shows that the user behavior who visits multiple Web site can be clustered. These clusters can be discovered by their algorithm based on the connectivity among Web sites [12]. Web Usage mining method includes sophisticated forms of analysis to find the common traversal paths through a Web site by using logs of large Web data in order to produce results that can be used in the design tasks. However, there are several preprocessing tasks that must be per-formed prior to applying data mining algorithms to the data collected from server logs. Study shows that several data preparation techniques can be used in order to identify unique users and user sessions. [13]. Another study reveals that parameter less replacement policies in Web proxies to handle client requests in an ISP environment and evaluating the performance of several existing policies can introduce Virtual Caches that can improve the performance of the cache for multiple metrics simultaneously [14].

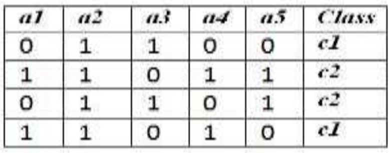
Another work with Web usage data shows how pattern discovery techniques such as clustering, association rule mining, and sequential pattern discovery can be leveraged effectively as an integrated part of a Web personalization system. [15].A Web cache replacement algorithms can be based on the Least Recently Used (LRU) idea by considering the number of references to Web objects as a critical parameter for the cache content replacement. And the algorithms are validated under Web cache traces provided by a major Squid proxy cache server installation environment. Cache and bytes hit rates are reported showing that the proposed cache replacement algorithms improve cache content [16].A case-based reasoning approach to discover user access patterns by mining the fuzzy association rules from the proxy log data where time duration of each user session is considered as one of the attributes of a web access case. A fuzzy index tree is used for fast matching of rules. An adaptation process is used to enhance system's performance [17].Another study present three methods of actionable Web log mining. The first is to mine a Web log using Hidden Markov models (HMV) for improving caching and prefetching of Web objects. The second one is to use the mined knowledge for building better, adaptive user interface. The last one is to apply Web query log knowledge to improving Web search [18].

A survey deals with the recent developments in Web log Mining area that is receiving increasing attention from the Data Mining community show the association rules method to find associations among web pages that frequently appear together in users’ sessions. Sequential Patterns and Clustering methods are also used to discover frequent subsequences among large amount of sequential data [19] .one study shows that Existing techniques of selecting pages cannot capture a user’s surfing patterns correctly. By using Weighted Association Rule (WAR) mining technique it is possible to classify pages of the user’s current interest and cache them to give faster net access. This approach captures both user’s habit and interest as compared to other approaches where emphasis is only on habit [20] .User navigation patterns discovery and analysis and privacy by Web mining can make special attention. By researching Web usage mining system, Web SIFT, it is possible to understand the methodology of how to apply data mining techniques to large Web data repositories in order to extract usage patterns [21] . However, Some Web usage mining phases called preprocessing, pattern discovery, and pattern analysis can provide a detailed taxonomy of the work web log mining, including research efforts as well as up-to-date survey of the existing work. By using Web SIFT system of a prototypical Web usage mining model can be formed [22].

In summary, it can be said that earlier work doesn't shows the extensive AR mining methods using Apriori Algorithm. The earlier used method shows only some discrete results in the proxy log mining area. Another common shortcoming of earlier works for proxy log mining is the lack of quality AR rule generation. In order to make AR rule mining productive for Knowing user access patterns the proposed method would be much more feasible.

The Problem Statement

Suppose ,in an examination E there are 10 questions Q ,say,"$(q1.q2,q3\dots\dots\dots.q10)$" .From E it is seen that majority of students S are passed in E with answering the Q no,for instance $q3,q4,q5$ .But on the other hand some S s are failed where they also attempt the Q no $q3,q4,q5$ .So, Here is the point why they failed .Is there anything wrong with the assessment. So, Here, it needs to be sort out that the S s who passed the E with answering Q no $q3,q4,q5$,are they really have the ability to pass and Ss who are failed with answering Q no $q3,q4,q5$,are they wrongfully failed. That's pattern needs to be check out. We fit this problem into our problem that clients who are getting access to information are they authorized for that information or not. Are they intruder? Let, I be a set of items such that $I = \{\square 1, \square 2, \dots\dots\dots\dots, \square n\}$ .An item set is a subset of items where it contains $(\square 1, \square 2, \quad \square 3)$from I . While on the other hand Transactions in the whole proxy log data set,D , and $(t1, t2, \dots\dots\dots\dots, tM)$, where t_i represents a transaction. Suppose an association rule has a form $X => Y$,where It is required that $X \subset I, Y \subset I$ and $X \bigcap \square Y = \square$The AR support for the rule $X => Y$ is $Sup(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{T} = \frac{number\ of\ transactions\ containing\ both\ X\&Y}{total\ number\ of\ transactions}$, σ represents support and the confidence of the rule is $Conf(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{number\ of\ transactions\ containing\ both\ X\&Y}{number\ of\ transactions\ containing\ X}$. Given a minimum confidence minconf , a rule is confident if $Conf(X \rightarrow Y) \geq$ minconf. From the proxy log Dataset D ,By Applying Aprioiri algorithm, we mine some rules R={r1,r2,….rm},where ri represent a rule. And the rule has a form $X => Y$,Where X is antecedent and Y is consequent. To formulate the association rule let there be 2 classes $c1$ and $c2$ .and A be a set of n attributes $(a1, a2\dots\dots.an)$. For example if the data set has a form like below, Where for Every Transaction t_i a class is either on $c1$ or $c2$.

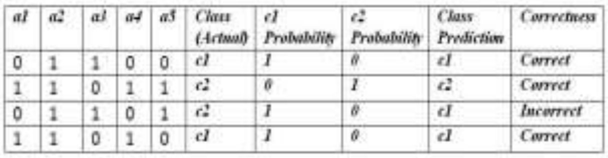


a1	a2	a3	a4	a5	Class
0	1	1	0	0	c1
1	1	0	1	1	c2
0	1	1	0	1	c2
1	1	0	1	0	c1

Fig 3.1: Data Set D in its initial Form

Proposed Solution

From the proxy log Dataset D , we calculate the 2 classes' probability and class prediction. Finally we get the correctness of the algorithm where the prediction is correct or not.



a1	a2	a3	a4	a5	Class (Actual)	c1 Probability	c2 Probability	Class Prediction	Correctness
0	1	1	0	0	c1	1	0	c1	Correct
1	1	0	1	1	c2	0	1	c2	Correct
0	1	1	0	1	c2	1	0	c1	Incorrect
1	1	0	1	0	c1	1	0	c1	Correct

Fig 4.1: Calculate Prediction and Correctness with both classes ($c1$ or $c2$) Probability.

From the generated rules with the class $c1$ **or** 2 , The Support $(Supp)$ Confidence $(Conf)$ of the each rule ri take into account. And then from the $Supp$ and $Conf$,The probability P is calculated ,where, $P = \frac{(Supp * Conf)}{100}$. [27]. If there is a match in the rule, the class probability (CP) for either $c1$ or $c2$ is averaged. Where $CP = \frac{\sum match\ rule\ probability}{n}$, where $n = number\ of\ matched\ rule$ .then comparing the 2 classes probability($c1$ or $c2$), which class probability is high is taken and set the class into the Class Prediction Column. After that the correctness of each transaction t_i is calculated. Where we find a match between Actual Class and Class Prediction Column we make it Correct in the Correctness column and if there is no match we make it incorrect in the Correctness column. Then the efficiency is calculated. Where the $efficiency(E) = \frac{Total\ Correctness}{Total\ Data}$. After that human judgment is performed to see whether the algorithm detection is right or wrong. and finally the proxy rules is updated. The proposed solutions are organized as follows:

Step1: Data Preparation
Step2: Rule Mining
Step3: Probability Calculation
Step4: Intruder Non Intruder Determination
Step5: Efficiency Calculation
Step6: Human Judging By Checking Dataset
Step7: Update Proxy rules from mined AR

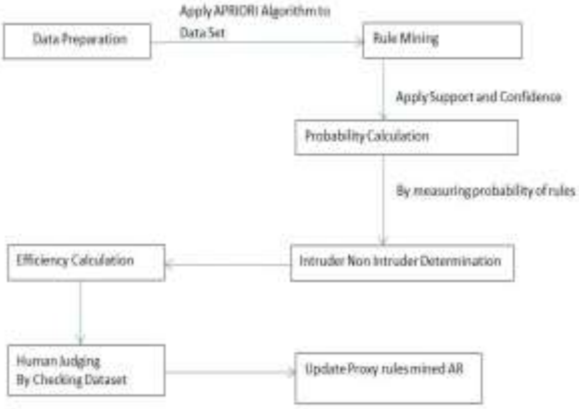
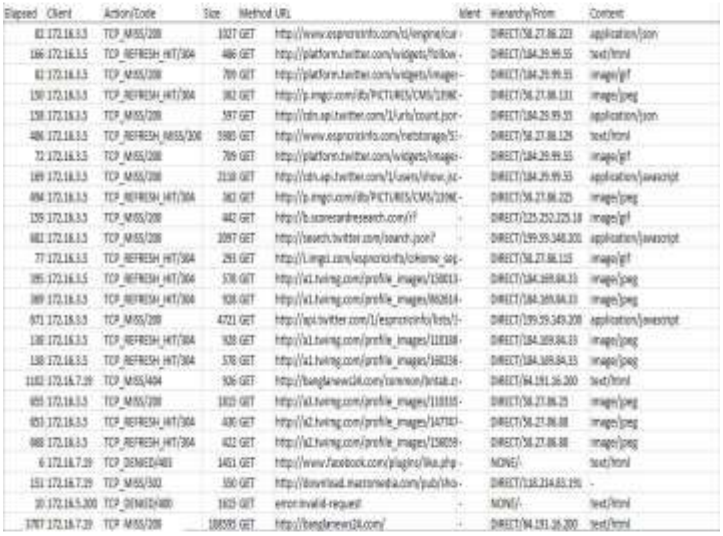

Data Preparation
Apply APRIORI Algorithm to Data Set
Rule Mining
Apply Support and Confidence
Probability Calculation
By measuring probability of rules
Intruder Non Intruder Determination
Efficiency Calculation
Human Judging By Checking Dataset
Update Proxy rules mined AR

Fig 4.2: Flow Chart Of proposed Solution

Step 1 Data Preparation

In our proposed solution the first step is to prepare the data to fit for processing into the Apriori algorithm to get proper rules.

At the initial stage, when the data is collected, the data is looked like below.



Elapsed	Client	Action/Code	Size	Method	URL	Ident	Hierarchy/From	Content
82	172.16.3.5	TCP_MISS/200	1027	GET	http://www.espncricinfo.com/ci/engine/cur	-	DIRECT/58.27.86.223	application/json
166	172.16.3.5	TCP_REFRESH_HIT/304	486	GET	http://platform.twitter.com/widgets/follow	-	DIRECT/184.29.99.55	text/html
82	172.16.3.5	TCP_MISS/200	709	GET	http://platform.twitter.com/widgets/images	-	DIRECT/184.29.99.55	image/gif
150	172.16.3.5	TCP_REFRESH_HIT/304	382	GET	http://p.imgci.com/db/PICTURES/CMS/1396	-	DIRECT/58.27.86.131	image/jpeg
158	172.16.3.5	TCP_MISS/200	597	GET	http://cdn.api.twitter.com/1/urls/count.json	-	DIRECT/184.29.99.55	application/json
486	172.16.3.5	TCP_REFRESH_MISS/200	5985	GET	http://www.espncricinfo.com/netstorage/5	-	DIRECT/58.27.86.128	text/html
72	172.16.3.5	TCP_MISS/200	709	GET	http://platform.twitter.com/widgets/images	-	DIRECT/184.29.99.55	image/gif
169	172.16.3.5	TCP_MISS/200	2138	GET	http://cdn.api.twitter.com/1/users/show.js	-	DIRECT/184.29.99.55	application/javascript
494	172.16.3.5	TCP_REFRESH_HIT/304	382	GET	http://p.imgci.com/db/PICTURES/CMS/1396	-	DIRECT/58.27.86.225	image/jpeg
139	172.16.3.5	TCP_MISS/200	442	GET	http://b.scorecardresearch.com/r?	-	DIRECT/125.252.225.18	image/gif
682	172.16.3.5	TCP_MISS/200	2097	GET	http://search.twitter.com/search.json?	-	DIRECT/199.59.148.201	application/javascript
77	172.16.3.5	TCP_REFRESH_HIT/304	293	GET	http://i.imgci.com/espncricinfo/ciHome_seg	-	DIRECT/58.27.86.115	image/gif
395	172.16.1.5	TCP_REFRESH_HIT/304	578	GET	http://a1.twimg.com/profile_images/15801	-	DIRECT/184.169.84.33	image/jpeg
389	172.16.3.5	TCP_REFRESH_HIT/304	928	GET	http://a1.twimg.com/profile_images/66261	-	DIRECT/184.169.84.33	image/jpeg
971	172.16.3.5	TCP_MISS/200	4721	GET	http://api.twitter.com/1/espncricinfo/lists/1	-	DIRECT/199.59.149.200	application/javascript
130	172.16.3.5	TCP_REFRESH_HIT/304	928	GET	http://a1.twimg.com/profile_images/11818	-	DIRECT/184.169.84.33	image/jpeg
138	172.16.3.5	TCP_REFRESH_HIT/304	578	GET	http://a1.twimg.com/profile_images/16023	-	DIRECT/184.169.84.33	image/jpeg
1102	172.16.7.39	TCP_MISS/404	926	GET	http://banglanews24.com/common/britab.c	-	DIRECT/64.191.26.200	text/html
655	172.16.3.5	TCP_MISS/200	1815	GET	http://a1.twimg.com/profile_images/11831	-	DIRECT/58.27.86.25	image/jpeg
651	172.16.3.5	TCP_REFRESH_HIT/304	430	GET	http://a2.twimg.com/profile_images/14718	-	DIRECT/58.27.86.88	image/jpeg
988	172.16.3.5	TCP_REFRESH_HIT/304	432	GET	http://a2.twimg.com/profile_images/15805	-	DIRECT/58.27.86.88	image/jpeg
6	172.16.7.39	TCP_DENIED/403	1451	GET	http://www.facebook.com/plugins/like.php	-	NONE/	text/html
151	172.16.7.39	TCP_MISS/503	350	GET	http://download.macromedia.com/pub/sho	-	DIRECT/118.214.83.191	-
10	172.16.5.200	TCP_DENIED/400	1615	GET	error:invalid-request	-	NONE/-	text/html
3707	172.16.7.39	TCP_MISS/200	188595	GET	http://banglanews24.com/	-	DIRECT/64.191.26.200	text/html

Fig 4.3: Main Dataset before Preparation

In the Data set Elapsed Column represents the time between the acception and closing of the client socket for specific request by client. This values is converted into too low, Low. Mid, High and Too High. by using Five number Summary. For Size and Intrusion Per IP Column same Approach is used. In the hierarchy/From Colum we set Direct/118.214.83.191 type data as Direct and None/- type data as none.

For the Content Colum we set 8 categories where App/zip file falls into Application category, Audio/Mpeg falls into Audio category, font/ttf falls into font category, image/bmp falls into image category, text/pdf falls into document category, Video/flv falls into Video Category, Where there is a blank those data falls into none category, and we have a others category as well where the content cannot be traced. The formatting is shown below.

Colum : Content	
Content Type	Category
App/zip	Application
Audio/Mpeg	Audio
font/ttf	Font
image/bmp	Image
text/pdf	Document
Video/flv	Video
Blank	None
Vague Data	Other

For the URL (Uniform Resource Locator) Column we set different categories. Those URL which fell into 117.79.92.35:443 ;this type of category we set them as others, URLs with msn.com,wikipedia.com,yahoo.com,google.com; are in the Search Engine category, URLs with facebook.com, Twitter.com; fell into Social network, URLs with youtube.com; treated as Entertainment, URLs with kespersky.com, Avg.com, Avast.com, Norton.com; are in the Antivirus category, URLs with yahoomail.com,gmail.com; are in the Mail category, URLs with cricinfo.com; are in the cricket category, URL which Shows error; this type of category we set them as Error, URLs with Blogspot.com; this type of category we set them as Blog, URL which fell into Yale.edu, adelade.edu.au etc ; this type of category we set them as Education, URLs with dictionary.cambridge.org, oxforddictionaries.com; are in the dictionariy category, URLs with bdjobs.com, prothom-alojobs.com; are in jobs category, URLs with nokia.com,mobilebd.com; are in the category mobile, URLs with amazon.com,ebay.com; are in online shopping, URLs with uiu.ac.bd; this type of category we set them as UIU(United International University), URLs with book related website; are in the Books category, URLs with Microsoft.com,Apple.com; are in Business. URL which seem unorganized; this type of category we set them as others. The formatting is shown below.

Colum : URL	
URL Type	Category
117.79.92.35:443	others
msn.com,wiki.com,yahoo.com,google.com	Search Engine
Facebook.com,Twitter.com	Social network
youtube. com	Entertainment
kesperskycom,Avgcom,Avast,Norton	Antivirus
yahoomailcom,gmail com	mail
cricinfo com	Cricket
URL which Shows error	error
Blogspot com	Blog
Yale.edu, adelade.edu.au	Education
dictionary.cambridge.org, oxforddictionaries.com	dictionariy
into bdjobs.com, prothom-alojobs.com	jobs
nokia.com,mobilebd.com	mobile
amazon.com,ebay.com	online shop
uiu.ac.bd	United International University
book related website	book
Microsoft.com,Apple.com	Business
URL which seem unorganized	others

Step 2 Rule Mining

From That prepared data we generate some rule by applying the Apriori algorithm. Here is how Apriori algorithm works.

Step 2.1 Apriori Algorithm steps

Apriori algorithm tries to find the frequent itemsets from the main superset. This algorithm generates candidate itemset of length k from the itemset of length (k-1).The Apriori algorithm states that if any length pattern ,suppose K,is not frequent in the whole dataset, its consecutive (K+1) pattern will never be frequent. The Apriori algorithm steps are as follows:

1. Scan the Database(D) to create Candiate Itemset $C1$
2. Then in Level 1 ($L1$) Get the itemset with greater support by eliminating the lower support.Where,$Support \geq minimum\ support\ threshold$.
3. From $L1$,make a combination of itemsets to get $C2$ (Candiate Itemset 2)
4. From $C2$, Generate $L2$ (Level 2) where , $Support \geq minimum\ support\ threshold$.
5. From $L2$ make a combination of itemsets to get $C3$ (Candiate Itemset 3)
6. From$C3$, Generate $L3$ (Level 3) where,$Support \geq minimum\ support\ threshold$.

Step 3 Probability Calculation

The Probability is calculated for every rule and set the probability values to the proxy log Database D . From the generated rules with the class $c1$ or , The Support $(Supp)$ Confidence $(Conf)$ of the each rule ri take into account. And then from the $Supp$ and $Conf$ The probability P is calculated ,where, $P = \frac{(Supp * Conf)}{100}$ .If there is a match in the rule, the class probability (CP) for either $c1$ or $c2$ is averaged. Where $CP = \frac{\sum match\ rule\ probability}{n}$, where $n = number\ of\ matched\ rule$ .then comparing the 2 classes probability ($c1$ or$c2$), which class probability is high is taken and set the class into the Class Prediction Column.

Step 4 Intruders Non Intruder Determination

The step looks for intruder non intruder in the proxy log Database D .which we are calling here as a prediction of getting whether the class is a intruder class or in non-intruder class. We get This Prediction by matching the 2 classes ($c1$ and $c2$) probability, where we get a class probability high by comparing those 2 classes ,we set the high class into the prediction column by treating them as intruder or non-intruder.

Step 5 Efficiency Calculation

Now the efficiency(E) of our proposed algorithm needs to be checked whether our algorithm is efficient or not and whether it is determining the intruder and non-intruder correctly. Where the efficiency is calculated with the following formula:

$$Proxy\,log\,efficiency(E) = \frac{Total\ Correctness}{Total\ Data\ in\ dataset}$$

If the $efficiency(E) \geq 80\%$ we treat it as an efficient algorithm.

Step 6 Human Judging By Checking Dataset

Rule mining are subjective to end user .Changing the threshold used by the rule mining can change mining of rule as well. For the professional use of those rules in various industries, Human judgment is very much necessary to clarify the rules .After the judgment if the found rules are not useful the threshold is changed to get the target set of rules that is required by the end user.

Step 7 Proxy Server Rule Update

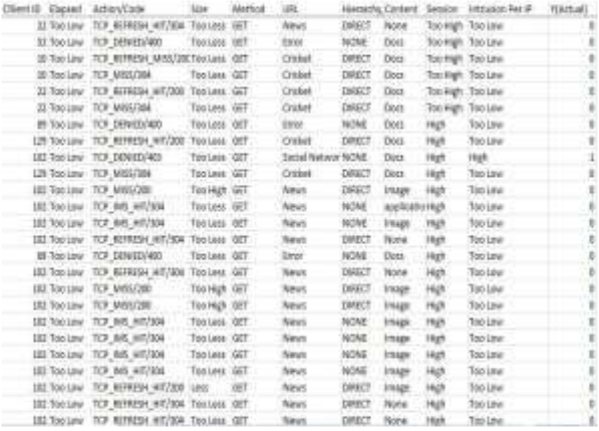
After getting the target rule that is required by the end user, we update the proxy server rule to get a good result. Where the updated rule can now detect more accurately that which client is intruder and which client is non-intruder.

5 Experimental Results

5.1 The Proxy log Data Set

The proxy log data set has been chosen for the evaluation of our Algorithm. It consists of more than 4 lakhs of samples. Each transaction in the proxy log Database represents a client's information who wants to access information. There are items like Elapsed, Action/Code, Size, Method, URL, Hierarchy/From, Content, Session and Intrusion Per IP where each items are categorical. For example Elapsed has a category of high, too high, medium, low too low and URL has Social Network, Search engine etc.

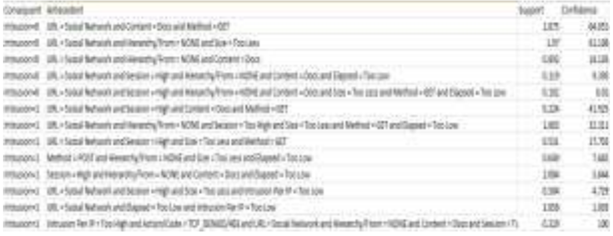
After The data preparation we get the proxy log Data set look like below. Where all the items like Elapsed, Action/code, size, Method, URL, Hierarchy/from, Content, Session, Intrusion per IP and The Actual classes are seen in the Data set.



Client ID	Elapsed	Action/Code	Size	Method	URL	Hierarchy	Content	Session	Intrusion Per IP	Y(Actual)
33	Too Low	TCP_REFRESH_HIT/304	Too Less	GET	News	DIRECT	None	Too High	Too Low	0
33	Too Low	TCP_DENIED/400	Too Less	GET	Error	NONE	Docs	Too High	Too Low	0
30	Too Low	TCP_REFRESH_MISS/200	Too Less	GET	Cricket	DIRECT	Docs	Too High	Too Low	0
30	Too Low	TCP_MISS/304	Too Less	GET	Cricket	DIRECT	Docs	Too High	Too Low	0
22	Too Low	TCP_REFRESH_HIT/200	Too Less	GET	Cricket	DIRECT	Docs	Too High	Too Low	0
22	Too Low	TCP_MISS/304	Too Less	GET	Cricket	DIRECT	Docs	Too High	Too Low	0
89	Too Low	TCP_DENIED/400	Too Less	GET	Error	NONE	Docs	High	Too Low	0
129	Too Low	TCP_REFRESH_HIT/200	Too Less	GET	Cricket	DIRECT	Docs	High	Too Low	0
182	Too Low	TCP_DENIED/403	Too Less	GET	Social Networ	NONE	Docs	High	High	1
129	Too Low	TCP_MISS/304	Too Less	GET	Cricket	DIRECT	Docs	High	Too Low	0
181	Too Low	TCP_MISS/200	Too High	GET	News	DIRECT	Image	High	Too Low	0
182	Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	applicatio	High	Too Low	0
182	Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	Image	High	Too Low	0
182	Too Low	TCP_REFRESH_HIT/304	Too Less	GET	News	DIRECT	None	High	Too Low	0
88	Too Low	TCP_DENIED/400	Too Less	GET	Error	NONE	Docs	High	Too Low	0
182	Too Low	TCP_REFRESH_HIT/304	Too Less	GET	News	DIRECT	None	High	Too Low	0
182	Too Low	TCP_MISS/200	Too High	GET	News	DIRECT	Image	High	Too Low	0
182	Too Low	TCP_MISS/200	Too High	GET	News	DIRECT	Image	High	Too Low	0
182	Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	Image	High	Too Low	0
182	Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	Image	High	Too Low	0
182	Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	Image	High	Too Low	0
182	Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	Image	High	Too Low	0
182	Too Low	TCP_REFRESH_HIT/200	Less	GET	News	DIRECT	Image	High	Too Low	0
182	Too Low	TCP_REFRESH_HIT/304	Too Less	GET	News	DIRECT	None	High	Too Low	0
182	Too Low	TCP_REFRESH_HIT/304	Too Less	GET	News	DIRECT	None	High	Too Low	0

Fig 5.1: Dataset after Preparation

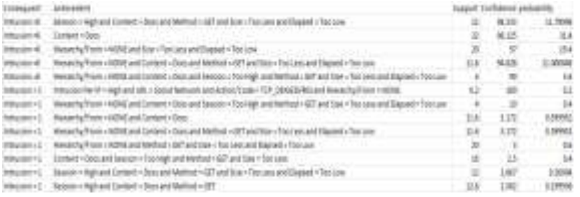
After completing the Data Preparation, Apriori algorithm is applied to The Prepared proxy log data to generate the desired rules with the minimum confidence and minimum support so that we can find our expected rules.



Consequent	Antecedent	Support	Confidence
Intrusion=0	URL = Social Network and Content = Docs and Method = GET	[illegible]	[illegible]
Intrusion=0	URL = Social Network and Hierarchy/From = NONE and Size = Too Less	[illegible]	[illegible]
Intrusion=0	URL = Social Network and Hierarchy/From = NONE and Content = Docs	[illegible]	[illegible]
Intrusion=0	URL = Social Network and Session = High and Hierarchy/From = NONE and Content = Docs and Elapsed = Too Low	[illegible]	[illegible]
Intrusion=0	URL = Social Network and Session = High and Hierarchy/From = NONE and Content = Docs and Size = Too Less and Method = GET and Elapsed = Too Low	[illegible]	[illegible]
Intrusion=1	URL = Social Network and Session = High and Content = Docs and Method = GET	[illegible]	[illegible]
Intrusion=1	URL = Social Network and Hierarchy/From = NONE and Session = Too High and Size = Too Less and Method = GET and Elapsed = Too Low	[illegible]	[illegible]
Intrusion=1	URL = Social Network and Session = High and Size = Too Less and Method = GET	[illegible]	[illegible]
Intrusion=1	Method = POST and Hierarchy/From = NONE and Size = Too Less and Elapsed = Too Low	[illegible]	[illegible]
Intrusion=1	Session = High and Hierarchy/From = NONE and Content = Docs and Elapsed = Too Low	[illegible]	[illegible]
Intrusion=1	URL = Social Network and Session = High and Size = Too Less and Intrusion Per IP = Too Low	[illegible]	[illegible]
Intrusion=1	URL = Social Network and Elapsed = Too Low and Intrusion Per IP = Too Low	[illegible]	[illegible]
Intrusion=1	Intrusion Per IP = Too High and Action/Code = TCP_DENIED/403 and URL = Social Network and Hierarchy/From = NONE and Content = Docs and Session = (T)	[illegible]	100

Fig 5.2: Example of Some Rules Found After Processing the Data through Apriori Algorithm

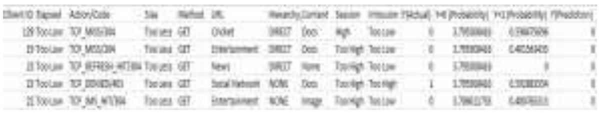
After The Rule mining step we need to calculate the probability of each rule



Consequent	Antecedent	Support	Confidence	probability
Intrusion=0	Session = High and Content = Docs and Method = GET and Size = Too Less and Elapsed = Too Low	12	[illegible]	[illegible]
Intrusion=0	Content = Docs	22	[illegible]	[illegible]
Intrusion=0	Hierarchy/From = NONE and Size = Too Less and Elapsed = Too Low	25	57	[illegible]
Intrusion=0	Hierarchy/From = NONE and Content = Docs and Method = GET and Size = Too Less and Elapsed = Too Low	11.6	[illegible]	[illegible]
Intrusion=0	Hierarchy/From = NONE and Content = Docs and Session = Too High and Method = GET and Size = Too Less and Elapsed = Too Low	4	99	[illegible]
Intrusion=1	Intrusion Per IP = High and URL = Social Network and Action/Code = TCP_DENIED/403 and Hierarchy/From = NONE	[illegible]	100	[illegible]
Intrusion=1	Hierarchy/From = NONE and Content = Docs and Session = Too High and Method = GET and Size = Too Less and Elapsed = Too Low	4	[illegible]	[illegible]
Intrusion=1	Hierarchy/From = NONE and Content = Docs	11.6	[illegible]	[illegible]
Intrusion=1	Hierarchy/From = NONE and Content = Docs and Method = GET and Size = Too Less and Elapsed = Too Low	[illegible]	[illegible]	[illegible]
Intrusion=1	Hierarchy/From = NONE and Method = GET and Size = Too Less and Elapsed = Too Low	20	1	[illegible]
Intrusion=1	Content = Docs and Session = Too High and Method = GET and Size = Too Less	18	2.5	[illegible]
Intrusion=1	Session = High and Content = Docs and Method = GET and Size = Too Less and Elapsed = Too Low	12	[illegible]	[illegible]
Intrusion=1	Session = High and Content = Docs and Method = GET	[illegible]	[illegible]	[illegible]

Fig 5.3: Figure showing the probability of some rules

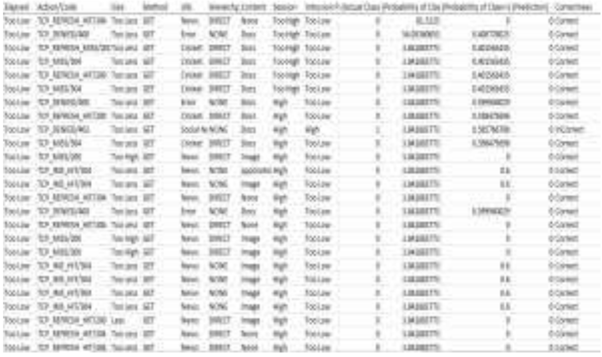
After that we are about to prediction of the intruder and non-intruder where 0 is set as non-intruder and 1 is set as intruder.



Client ID	Elapsed	Action/Code	Size	Method	URL	Hierarchy	Content	Session	Intrusion	Y(Actual)	Y=0 (Probability)	Y=1 (Probability)	Y(Prediction)
129	Too Low	TCP_MISS/304	Too Less	GET	Cricket	DIRECT	Docs	High	Too Low	0	[illegible]	[illegible]	0
23	Too Low	TCP_MISS/304	Too Less	GET	Entertainment	DIRECT	Docs	Too High	Too Low	0	[illegible]	[illegible]	0
23	Too Low	TCP_REFRESH_HIT/304	Too Less	GET	News	DIRECT	None	Too High	Too Low	0	[illegible]	0	0
23	Too Low	TCP_DENIED/403	Too Less	GET	Social Network	NONE	Docs	Too High	Too High	1	[illegible]	[illegible]	[illegible]
21	Too Low	TCP_IMS_HIT/304	Too Less	GET	Entertainment	NONE	Image	Too High	Too Low	0	[illegible]	[illegible]	0

Fig 5.4 prediction of the intruder and non-intruder

Then The efficiency step comes for the proxy log efficiency. Where the efficiency is calculated with the total correct data as a numerator and total transaction amount in the data set as a Denominator. From our experiment, from the proxy log data set, it is found that we got efficiency more than 80%.



Elapsed	Action/Code	Size	Method	URL	Hierarchy	Content	Session	Intrusion Per IP	Actual Class	Probability of Class 0	Probability of Class 1	(Prediction)	Correctness
Too Low	TCP_REFRESH_HIT/304	Too Less	GET	News	DIRECT	None	Too High	Too Low	0	[illegible]	0	0	Correct
Too Low	TCP_DENIED/400	Too Less	GET	Error	NONE	Docs	Too High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_REFRESH_MISS/200	Too Less	GET	Cricket	DIRECT	Docs	Too High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_MISS/304	Too Less	GET	Cricket	DIRECT	Docs	Too High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_REFRESH_HIT/200	Too Less	GET	Cricket	DIRECT	Docs	Too High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_MISS/304	Too Less	GET	Cricket	DIRECT	Docs	Too High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_DENIED/400	Too Less	GET	Error	NONE	Docs	High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_REFRESH_HIT/200	Too Less	GET	Cricket	DIRECT	Docs	High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_DENIED/403	Too Less	GET	Social Ne	NONE	Docs	High	High	1	[illegible]	[illegible]	0	InCorrect
Too Low	TCP_MISS/304	Too Less	GET	Cricket	DIRECT	Docs	High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_MISS/200	Too High	GET	News	DIRECT	Image	High	Too Low	0	[illegible]	0	0	Correct
Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	applicatio	High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	Image	High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_REFRESH_HIT/304	Too Less	GET	News	DIRECT	None	High	Too Low	0	[illegible]	0	0	Correct
Too Low	TCP_DENIED/400	Too Less	GET	Error	NONE	Docs	High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_REFRESH_HIT/304	Too Less	GET	News	DIRECT	None	High	Too Low	0	[illegible]	0	0	Correct
Too Low	TCP_MISS/200	Too High	GET	News	DIRECT	Image	High	Too Low	0	[illegible]	0	0	Correct
Too Low	TCP_MISS/200	Too High	GET	News	DIRECT	Image	High	Too Low	0	[illegible]	0	0	Correct
Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	Image	High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	Image	High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	Image	High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_IMS_HIT/304	Too Less	GET	News	NONE	Image	High	Too Low	0	[illegible]	[illegible]	0	Correct
Too Low	TCP_REFRESH_HIT/200	Less	GET	News	DIRECT	Image	High	Too Low	0	[illegible]	0	0	Correct
Too Low	TCP_REFRESH_HIT/304	Too Less	GET	News	DIRECT	None	High	Too Low	0	[illegible]	0	0	Correct
Too Low	TCP_REFRESH_HIT/304	Too Less	GET	News	DIRECT	None	High	Too Low	0	[illegible]	0	0	Correct

Fig 5.5: The Correctness of each transaction is calculated.

From our experiment we found a total correctness of 340792, whereas the total data in the data set is 402987.

So,

$$Proxy\,log\,efficiency(E) = \frac{340792}{402987} = .847 \text{ (approx.)} = 84.7\%$$

Though we got an efficiency of more than 80%, we again perform the human judging process to check whether our extracted information is correct or not. The finally we update the proxy server rule To detect the intruder non-intruder efficiently.

Conclusion and Future work

Mined Rules can play a fundamental rule. Our Paper uses the Apriori Algorithm to mine desired rule from the proxy log data set. We set extra class information to mark the transactions as intruder or non-intruder. Our method has an efficiency of over 80% to mark the intruder non intruder class. Our method is a step by step process where at first we make our data preparation to make the data fit for the processing after that we extract Rules from the data set using Apriori algorithm. Then the probability of each rule is calculated. By measuring the probability of rules we can determine intruder non intruder. After that the efficiency of the algorithm is calculated by dividing the total correct data with the Total data. Then Human Judging is performed to check whether the algorithm is working properly or not. Finally from the mined association rule we update the Proxy rules to make the intruder non intruder detection efficient. In future we are planning to work with the ACCESS Log to experiment the results of which type of contents accessed by clients are intruder and non-intruder.

References:

[1] Pallis G, Vakali A. (2006) 'Insight and perspectives for content delivery networks',CommunACM (CACM);49(1):101–6.

[2]Pallis G, Vakali A, Pokorný J. (2008) 'A clustering-based prefetching scheme on a Web cache environment', Computers & Electrical Engineering, 34(4): 309-323.

[3]Agarwal, R., Ghosh, S., Imielinski, T., Lyer, B. and Swami, A.(1992) 'An interval classifier for database mining applications', International Conference on Very LargeDatabases (VLDB), Vancouver, Canada, pp.560–573.

[4]Han, J. and Kamber, M. (2001) ' Data Mining: Concepts and Techniques', Morgan Kaufmann, San Francisco,CA ,pp.279–333,

[5] Rahman M, Kabir F.andSiddiky F.A. 'Rules mining from multi-layered neural networks',Int. J. Computational Systems Engineering, Vol. 1, No. 1, 2012

[6].Fisher D, Saksena G. (2003) 'Link prefetching in Mozilla: a server-driven approach'. In Proceedings of the WCW;.

[7].G. Pallis, A. Vakali, J. Pokorny. (2008)'A Clustering-based Approach for Short-term Prefetching on a Web Cache Environment', Computers & Electrical Engineering Journal, Elsevier, 34(4): 309-323.

[8]Baoyao Zhou, Siu Cheung Hui, Kuiyu Chang (2004) 'A Formal Concept Analysis Approach for Web Usage Mining', Intelligent Information Processing: 437-440

[9]KoroshGolnabi, Richard Min, Latif Khan and Ehab Al-Shaer. (2006), 'Analysis of Firewall Policy Rule Using Data Mining Techniques', 10th IEEE/IFIP Network Operations and Management Symposium (NOMS 2006), April.

[10]D. Brauckhoff, X. Dimitropoulos, A. Wagner, and K. Salamatian. (2011) 'Anomaly extraction in backbone networks using association rules'. IEEE/ACM Transactions on Networking (under submission).

[11]C. C. Aggarwal, J. L. Wolf, P. S. Yu. (1999) 'Caching on the Worldwide Web'. IEEE Transactions on Knowledge and Data Engineering, Vol 11, No 1, January .

[12]Wenwu Lou, Guimei Liu, Hongjun Lu, Qiang Yang.(2002) 'Cut-and-Pick Transactions for Proxy Log Mining'. EDBT, pp:88-105

[13]V. Sathiyamoorthi and V. MuraliBhaskaran (2009) 'Data Preparation Techniques for Web Usage Mining in World Wide Web-An Approach', November.

[14]Martin F. Arlitt, Ludmila Cherkasova, John Dilley, Rich Friedrich, Tai Jin (2000) 'Evaluating content management techniques for Web proxy caches'. SIGMETRICS Performance Evaluation Review 27(4): 3-11

[15]Munindar P. Singh (ed.), CRC Press. 2005 'Web Usage Mining and Personalization. In Practical Handbook of Internet Computing,

[16]Vakali A (2000) 'LRU-based Algorithms for Web Cache Replacement',Proceeding EC-WEB '00 Proceedings of the First International Conference on Electronic Commerce and Web Technologies ,Springer-Verlag London, UK ©2000,ISBN:3-540-67981-2

[17]Cody K. P. Wong, Simon C. K. Shiu and Sankar K. Pal (2001), 'Mining Fuzzy Association Rules for Web Access Case Adaptation', Proceedings of Soft Computing in Case-Based Reasoning Workshop, in conjunction with the 4th International Conference in Case-Based Reasoning, Vancouver, Canada, 30-July to 2-August , pp.213-220

[18] Qiang Yang, Charles X. Ling and JianfengGao. (2004) 'Mining web logs for actionable knowledge'. Book chapter in NingZhong and Jiming Liu, editors, Intelligent Technologies for Information Analysis. Springer.

[19] Federico Michele Facca, Pier Luca Lanzi (2003) 'Recent Developments in Web Usage Mining Research', DaWaK: 140-150

[20] AbhinavSrivastava, AbhijitBhosale, Shamik Sural ' Speeding Up Web Access Using Weighted Association Rules' , PReMI2005: 660-665

[21] Y Wang (2000) 'Web mining and knowledge discovery of usage patterns', CS748T Project (Part I) Feb,

[22]Jaideep Srivastava, Robert Cooley, MukundDeshpande, Pang-Ning Tan (2000) 'Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data', SIGKDD Explorations 1(2): 12-23

[23]http://en.wikipedia.org/wiki/data_mining

[24]Jiawei Han and MichelineKamber (2006) ' Data Mining: Concepts and Techniques', Elsevier, ISBN 1558609016. This lecture notes is based on materials in chapter 5. 3, 27

[25]Lauri Lahti,'Apriori algorithm', Seminar of Popular Algorithms in Data Mining and Machine Learning, TKK Presentation 12.3.2008

[26]http://www.purplemath.com/modules/boxwhisk2.htm

[27] Bing Liu, Yiming Ma, Ching-Kian Wong, and Philip S. Yu. "Scoring the Data Using Association Rules." Applied Intelligence, Vol 18, No. 2, 119-135, 2003