International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:7, 2013

Improving the Performance of Proxy Server by Using Data Mining Technique

P. Jomsri

***Abstract*—**Currently, web usage make a huge data from a lot of user attention. In general, proxy server is a system to support web usage from user and can manage system by using hit rates. This research tries to improve hit rates in proxy system by applying data mining technique. The data set are collected from proxy servers in the university and are investigated relationship based on several features. The model is used to predict the future access websites. Association rule technique is applied to get the relation among *Date, Time, Main Group web, Sub Group web*, and *Domain name* for created model. The results showed that this technique can predict web content for the next day, moreover the future accesses of websites increased from 38.15% to 85.57 %.

This model can predict web page access which tends to increase the efficient of proxy servers as a result. In additional, the performance of internet access will be improved and help to reduce traffic in networks.

***Keywords*—**Association rule, proxy server, data mining.

I. Introduction

AT present, internet technology becomes important application for searching information as WWW. This technology reduces a limitation of time and place and they were used in academic organization which user can use such a system to survey academic information. Therefore the large information of internet traffic is increased. Many organizations use "Proxy server" and utilize the process of caching, storing data temporarily to help decrease the amount of traffic at any given time. Proxy servers are designed with three goals: decrease network traffic, reduce user (client) perceived lag, and reduce loads on the origin servers [1], [2].

Some problems appear from the fact which is a limited amount of hard disk space that proxies can save log files or web contents. The samples of decision-making problems are cache placement, cache consistency, and cache replacement. Main propose of this paper is the cache replacement problem, the process of evicting (page out) log files in the cache for new log files. However, log files that were evicted, sometime will be reused again. Therefore, proxy server will lost time to get the original web. This process has an effect on efficiency of proxy server which can measure by using the Hit Rate.

This paper tries to improve such a hit rate using data mining technique to predict web page which will call in the future.

The paper is organized as follows. Section II discusses related works. The framework of this paper is described in Section III. The experimental setting and association rule analysis explained in Section IV. The results and discussions from the experiments are presented in Section V. Finally, the conclusion and future work are given in Section VI.

P. Jomsri is with the Faculty of science and technology, Suan Sunandha Rajabhat University, Dusit, Bangkok 10300 Thailand (phone: +6602-160-1143; e-mail: pijitra.jo@ssru.ac.th, pijitra_jom@hotmail.com).

II. Related Work

This section contains two parts: first is background of Proxy server and second relate with Squid content.

A. Proxy Server

Proxy server is a server (a computer system or an application) that acts as an intermediary for the requests from clients seeking resources from other servers. A client connects to the proxy server, requesting some services, such as a file, connection, web page as follow in Fig. 1.

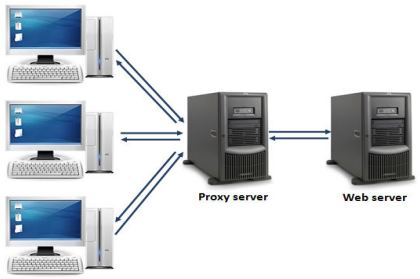

Fig. 1 A Framework for proxy server system

There are many replacement strategies to consider when designing a proxy server. The most commonly known cache replacement strategies are Least Frequently Used (LFU) and Least Recently Used (LRU). There had been no survey of known web cache replacement strategies before year 2003. However, Podlipnig et al. [3] have done well not only list well-known strategies, but also categorize the strategies into five groups: Frequency Based, Recency Based, Frequency/Recency Based, Function Based, and Randomized.

Many researchers are interested in web caching and its associated problems Luo et al. [4] focused on making proxy caching work for database-backed Web sites.-Houtzager et al. [5] proposed an evolutionary approach to find an optimal solution to the web proxy cache placement problem, while Aguilar and Leis [6], for example, addressed the replacement problem. Fagni et al. [7] proposed a static dynamic cache. They store the most popular web queries in a static, read-only portion of a cache. The remaining cache entries are dynamic and store other queries that cannot be satisfied using the static cache. The dynamic cache is managed by any replacement policy. In 2009, Kaya et al. [8] devised an admission-control

policy to screen documents based on a mathematical expression that is function of average delay per request. They use this policy to identify cacheable and non-cacheable documents and then use LRU for cache replacement.

Many cache replacement policies use artificial intelligence techniques for decision making. Sabeghi et al. [9] and Calzarossa [10] use fuzzy logic for improving cache replacement decisions. A survey on applications of neural networks and evolutionary techniques in web caching can be found in [11]. Khalid proposed a neural network-based cache replacement scheme called KORA.

Jomsri et al. [12] select log file factor such as date time and only main group web for use association rule to predict web content. Furthermore, they found that the efficiency of proxy can improve by using association rule.

B. Squid

Squid is a caching proxy for the Web supporting HTTP, HTTPS, FTP, and more. It reduces bandwidth and improves response times by caching and reusing frequently-requested web pages. Squid has extensive access controls and makes a great server accelerator. It runs on most available operating systems, including Windows and is licensed under the GNU GPL.

Thousands of web-sites around the Internet use Squid to drastically increase their content delivery. Squid can reduce your server load and improve delivery speeds to clients. Squid can also be used to deliver content from around the world - copying only the content being used, rather than inefficiently copying everything. Finally, Squid's advanced content routing configuration allows you to build content clusters to route and load balance requests via a variety of web servers.

Through Squid's configuration files and easy compilation steps, one can control almost any aspect of the cache including the replacement strategy, and how the proxy logs requests, such as keeping a log of the requests that Squid automatically saves to the hard disk. In the early 1990s, the Squid project came about from a fork of the Harvest Cache Daemon. The other fork existing today is Netapp's Netcache [13], which is a proprietary software package.

Therefore, an *access.log* entry usually consists of (at least) 10 columns separated by one or more spaces:

- **Time** A Unix timestamp as UTC seconds with a millisecond resolution.
- **Duration** The elapsed time considers how many milliseconds the transaction busied the cache. It differs in interpretation between TCP and UDP.
- **Client address** The IP address of the requesting instance, the client IP address.
- **Result codes** This column is made up of two entries separated by a slash. This column encodes the transaction result.
- **Bytes** The size is the amount of data delivered to the client. Mind that this does not constitute the net object size, as headers are also counted. Also, failed requests may deliver an error page, the size of which is also logged here.
- **Request method** The request method to obtain an object. Please refer to section request-methods for available methods.
- **URL** This column contains the URL requested. Please note that the log file may contain whitespace for the URI. The default configuration for *uri_whitespace* denies or truncates whitespace, though.
- **Rfc931** The eighth column may contain the ident lookups for the requesting client.
- **Hierarchy code** The hierarchy information consists of three items such as any hierarchy tag may be prefixed with *TIMEOUT*, code that explains how the request was handled, the IP address or hostname where the request (if a miss) was forwarded. For requests sent to origin servers, this is the origin server's IP address.
- **Type** The content type of the object as seen in the HTTP reply header.

This paper uses different views to improve the performance of proxy server by focusing in the reliability. This paper extends the method of association rule that is data mining technique to survey the relation of log content.

III. Framework for Predict Web Content by Using Association Rule

A framework to predict web content of proxy server is in Fig. 3. The framework mechanism is designed in six steps:

1) *Data Integration:* This process collects data from 2 sources. First is access log from proxy server , example of access log is shown in Fig. 2 and second is web content such as Group web and sub Group web which collect from www.mesook.com as show in Table I .
2) *Data cleaning:* In this process make from log file of proxy server by select the important data such as URL.
3) *Data conversion:* This is a conversion of the log file and web content data into the format needed by mining algorithms.

```
1128531640.658    148 202.44.135.35 TCP_REFRESH_HIT/200 504 GET
http://www.sanook.com/menu/images/sm2nbg.gif -
DIRECT/203.107.136.7 image/gif
1128531640.704    211 202.44.135.35 TCP_MISS/200 4809 GET
http://www.sanook.com/menu/nav.php -
TIMEOUT_DIRECT/203.107.136.7 text/html
1128531647.627    116 172.27.7.35 TCP_MISS/200 338 GET
http://truehits2.gits.net.th/biggen.php? - DIRECT/164.115.2.146
image/jpeg
1128531650.101    119 172.27.7.35 TCP_MISS/200 338 GET
http://truehits2.gits.net.th/biggen.php? - DIRECT/164.115.2.146
image/jpeg
```

Fig. 2 Example of access log

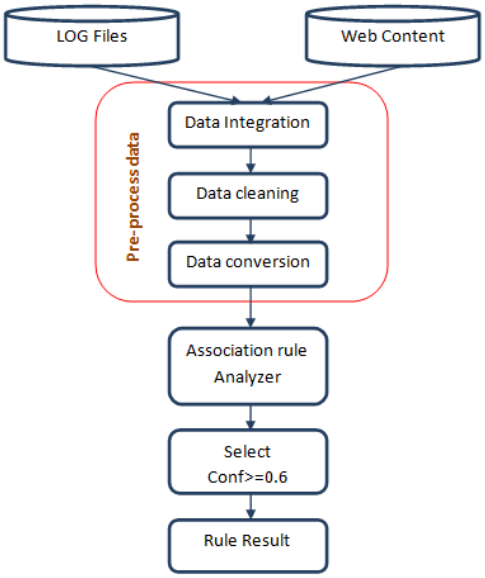

Fig. 3 A Framework for prediction web content by using association rule

TABLE I
THE INFORMATION OF WEB CONTENT

Group web	*Total Number*
sports	7
Bank and Financial institution	10
News and media	119
entertainment	698
internet	143
Knowledge and information	59
Government organization	68
academic	42
Personal social and culture	5
computer	405
Not be categorized	378

4) *Association rule*: This step is creating association rule model. The relation during time, main group web, sub group web and web is analyzed. This technique is recommending base on similarity and is describe in Section IV.
5) *Threshold:* We conduct this informal evaluation by setting the parameter α = 0.6 as the threshold value.
6) *Result:* The last step of the framework is the result. This step conclude the rule that the value of parameter α more than 0.6. Furthermore, the statistic is used in process to test the difference.

IV. ASSOCIATION RULE

Association rule discovery is a popular data mining method and well researched method for discovering interesting relations between variables in large databases. Many researches lead data mining and association rule for analyzes and increase efficiency in searching result [12], [14].

This paper analyzed basic data relation by using association rule discovery from access log and web content to improve proxy server. Researcher explored association of a set *time, main group web, sub group web, and domain name*. This research expects that the main group web, sub group web can help to predict web page and improve the performance of proxy server. Moreover this experiment uses data set from proxy at 902,420 records.

A. Association Rule Discovery

The rules tab in the form of $X \rightarrow Y$ is applied for extracting rules. Where *time*, *main group web, sub group web*, and *domain name* are disjoint item sets of access log. For each rule of the form $X \rightarrow Y$, researcher defines the *supp* and *conf* as the *support* and *confidence* as follows.

$$conf(X,Y) = \frac{count(X,Y)}{count(X)} \quad (1)$$

such as *conf* (*time*, *main group web, sub group web*, *web*)

$$= \frac{count(time, main\ group\ web, sub\ group\ web, domain\ name)}{count(time, main\ group\ web, sub\ group\ web\ )} \quad (2)$$

$$sup(X,Y) = \frac{count(X,Y)}{count(All)} \quad (3)$$

such as *sup*(*time*, *main group web, sub group web*, *domain name*)

$$= \frac{count(time, main\ group\ web, sub\ group\ web, domain\ name)}{count(All\ )} \quad (4)$$

0.00AM–1.00 AM, computer, game→www.facebook.com (5)

An example, the relation of web content was showed in (5). The rule explains that during 0.00AM–1.00 AM, the website in computer group and game subgroup which www.facebook.com was request by user.

The equation shows the rules for predicting the website during 0.00AM–1.00 AM. Confidence and support value are used for rule selections. Because plenty of rules are generated, some simple concerns in rule selections include:

1) Select the rule with maximum confidence.
2) Select the rule with maximum support if confidence value is equal.
3) Select the rule that happens first when confidence and support values are equal.

Table II shows the rule explanation:

- Support of *time, main group web, sub group web→ domain name* is the probability that web has both *time, main group web, sub group web* and *domain name*
- Confidence of *time, main group web, sub group web→ domain name* is probability that *web* is given the content appear in *time, main group web, sub group web*

TABLE II
EXAMPLES OF RELATION MODELS OF WEB CONTENT AT 0.00AM–1.00 AM WITH CONFIDENCE AND SUPPORT VALUES

Rule no.	*rule*	*Conf (%)*	*Sup (%)*
1	sports, Golf → www.tga.or.th	100%	0.19%
2	Bank and Financial institution,Bank →www.bangkokbank.com	52%	2.49%
3	entertainment,catoon → www.yimkrim.com	100%	12.18%
4	internet , download → www.siamware.com	100%	6.77%
5	Knowledge and information,forecast weather →www.gts.tmd.go.th	94.03%	1.42%
6	Knowledge and information,check lot → www.glo.or.th	9.50%	5.61%
7	Personal social and culture,board news →board.dserver.org	32%	0.29%
8	Personal social and culture, board news →geocities.com	2.3%	0.05%
9	computer, all computer →www.bcoms.net	80%	0.98%
10	computer, Macintosh →www.geocities.com	21%	0.78%
11	academic, teaching →www.chula.ac.th	97.78%	5.31%
12	academic,Institute of high education →www.ssru.ac.th	61.53%	0.78%
13	News and media ,program TV →www.ubcaf.com	21%	0.69%
14	News and media,television station→www.ch7.com	100%	0.45%
15	News and media , newspaper→ www.dailynews.co.th	100%	1.58%
16	News and media , newspaper → http://tvThaiOnline.th.gs	7.76%	4.03%
17	Government organization,Ministry of finance → www.mof.go.th	46.41%	32.03%
18	Not be categorized, Not be categorized →kapook.com	7.76%	8.25%
19	Not be categorized, Not be categorized → diary.sanook.com	20.77%	15.25%
20	Not be categorized, Not be categorized → community.ubcaf.com	58%	0.14%

B. Result of Association Rule Discovery

Table II shows total association prediction model for web content of proxy server with confidence and support values. Twenty rules were creating based on the frequency request.

V. RESULT AND DISCUSSION

This section separate into two parts: first is results from the experiment and the second is the discussion.

A. Results

The rules which have confidence value more than 60% were chosen and are shown in Table III. The strength rules were hold such as main group is sports and subgroup is golf at 0.00 AM–1.00 AM will request www.tga.or.th. This rule has Confidence value equal to 100% and Support value equal to 0.19%. Therefore, the relationship of this rule may help to improve the performance of proxy server.

TABLE III
CONFIDENCE OF ASSOCIATION RULE AT 0.00AM–1.00 AM WHERE $\alpha = 60\%$

Rule	*Conf (%)*	*Sup (%)*
sports, Golf → www.tga.or.th	100%	0.19%
entertainment,catoon → www.yimkrim.com	100%	12.18%
internet , download → www.siamware.com	100%	6.77%
Knowledge and information,forecast weather →www.gts.tmd.go.th	94.03%	1.42%
computer, all computer →www.bcoms.net	80%	0.98%
academic, teaching →www.chula.ac.th	97.78%	5.31%
academic,Institute of high education →www.ssru.ac.th	61.53%	0.78%
News and media,television station→www.ch7.com	100%	0.45%
News and media , newspaper→ www.dailynews.co.th	100%	1.58%

In addition, the accuracy of the relation was computed by dividing the data set into training set and testing set. The results of the test model is shown in the T and F (where T is the test model accuracy, F is the model with the error). Table IV shows the result of the accuracy.

Rule No. 1 can describe that the confidence value of training set is 82% and confidence value of test set is 100%.Therefore, the first order of accuracy equal to T.

Rule No. 5 describes that the confidence value of training set is 72.44% and confidence value of test set is 23.24%.Therefore, the first order of accuracy equal to F.

TABLE IV
THE ACCURACY OF CONFIDENCE TRAINING VALUE AND CONFIDENCE TESTING VALUE AT 0.00AM–1.00 AM

Rule No.	*Rule*	*Conf (%) Train*	*Conf (%) Test*	*Result*
1	sports, Golf → www.tga.or.th	82.00	100.00	T
2	entertainment,catoon → www.yimkrim.com	100.00	100.00	T
3	internet , download → www.siamware.com	95.43	97.13	T
4	Knowledge and information,forecast weather →www.gts.tmd.go.th	40.01	45.45	T
5	computer, all computer →www.bcoms.net	72.44	23.24	F
6	academic, teaching →www.chula.ac.th	100.00	100.00	T
7	academic,Institute of high education →www.ssru.ac.th	80.00	100.00	T
8	News and media,television station →www.ch7.com	75.00	100.00	T
9	News and media , newspaper → www.dailynews.co.th	98.98	99.04	T
10	News and media , newspaper → www.dailynews.co.th	82.00	100.00	T

To test the mean difference of the Model between *time, main group web, sub group web→ domain name* so called *GSRModel* and previous model so called *GRModel*, a paired-sample T test is employed. Assume that the sample comes

from populations that are approximately normal with equal variances. Level of significance is set to 0.05 (α=0.05). The results can be summarized as follows:

TABLE V
PAIRED-SAMPLE T TEST

	Pair differences			
Pair	*Mean*	*Std. D.*	*Std.error Mean*	*Sig (2-Tailed)*
GRModel -GSRModel	.092	.410	.023	.001

The statistical testing result from Table V indicates that there is a significant difference in the confidence values of the *GSRModel* and *GRModel* at α=0.05. In other words, the mean scores of confidence values of *GSRModel* and *GRModel* are not the same.

B. Discussion

This paper presents data mining techniques for improving proxy server based on the relation of web content. In the association rule step, the *support* and *confidence value* were used to determine the web relation. The performance of system which includes association rule technique by using *time, main group web, sub group web, and domain name* is more accuracy than the original system**.** Therefore, the relation of web content has a potential and can be used the technique for improving the performance of proxy server.

VI. CONCLUSION AND FUTURE WORK

The use of data mining technique can analyzes the relationship for predict web page. Particularly, the useful of prediction web applications can increase Hit Rate of proxy server.

In fact, during our study, it is becoming clear that only relying on one method to predict the preference. Furthermore, in this mechanism only one-to-one association rule is considers such as *time, main group web, sub group web* → *domain name*. This assumption is to simplify the problem.

In addition, This paper preliminary analysis a relation of the web content that appear the same group by using association rule discovery. Result of preliminary analysis of there some rule is interesting and can bring to improve performance proxy server.

In the future, researcher plan to use this information to enhance the analysis for improving proxy server.

ACKNOWLEDGMENT

The author would like to thank Suan Sunadha Rajabhat University for scholarship and support.

REFERENCES

[1] S .Podlipnig, L. Boszormenyi, "A survey of web cache replacement strategies," *ACM Comput Surv* ,vol.35(4), pp.374–398, 2003.
[2] B. Davison, "A web caching primer," *IEEE Internet Computing* vol.5(4), pp.38–45,2001.
[3] P. Jomsri, P. Tantasanawong, "Hit Rate Improvement in Proxy System using Data Mining Technique," *in Proc. National Conference on Information Technology*, Bangkok, 2006 .
[4] L. Qiong , N. F. Jeffrey, X. Wenwei, "Form-based proxy caching for database backed web sites: keywords and functions," VLDB J , vol.17(3), pp. 489–513 ,2008.
[5] G. Houtzager, C. Jacob, C. Williamson, "An evolutionary approach to optimal web proxy cache placement," *in proc IEEE Congr Evolut Comput ,2006.*
[6] J. Aguilar, EL. Leis ,"A coherence-replacement protocol for web proxy cache systems," *Int J Comput Appl* , vol.28(1), pp. 12–18, 2006.
[7] T. Fagni , R. Perego, S. Silverti, and S. Orlando,"Boosting the performance of web search engines: caching and prefetching query results by exploiting historical usage data," ACM Transactions on Information Systems, Vol. 24(1), pp. 51–78, 2006.
[8] C.C. Kaya, G. Zhang, Y. Tan, and V.S. Mookerjee , "An admission-control technique for delay reduction in proxy caching," *Decision Support Systems,* vol. 46(2), pp.594–603, 2009.
[9] M. Sabegi, and M. Yaghmaee, "Using fuzzy logic to improve cache replacement decisions," *IJCSNS International Journal of Computer Science and Network Security*, vol.6(3A), 2006.
[10] M.C. Calzarossa, and G. Valli, "A fuzzy Algorithm for web caching," *Simulation Series Journal*, vol. 35(4), pp. 630–636 ,2003.
[11] P. Venketesh, and R. Venkatesan, "A survey on applications of neural networks and evolutionary techniques in web caching," *IETE Tech Rev*, vol. 26(3),pp. 171–180, 2009.
[12] H. Khalid, "A new cache replacement scheme based on back propagation neural networks," *ACM SIGARCH Comput Archit News*, vol. 25(1), pp. 27–33, 1997.
[13] What is Squid?'', Available at http://www.squid-cache.org/Intro/
[14] X. Chen, and Y. Wu. "Personalized Knowledge Discovery: Mining Novel Association Rules from Text," Available: http://www.siam.org/meetings/sdm06/proceedings/067chenx.pdf