MINING DIBRUGARH UNIVERSITY DISTANCE EDUCATION WEBSITE AND BAYESIAN APPROACH FOR WEB PROXY CACHE MANAGEMENT

[1]SADIQ HUSSAIN, [2]JITEN HAZARIKA, [3]G.C. HAZARIKA

[1]System Administrator, Dibrugarh University, Dibrugarh, Assam, India

[2] Professor, Department of Statistics, Dibrugarh University, Assam, India

[3] Professor, Department of Mathematics, Dibrugarh University, Assam, India

E-mail: [1]sadiq@dibru.ac.in, [2]jiten_stats@dibru.ac.in, [3]gchazarika@dibru.ac.in

ABSTRACT

Web proxy cache management is a way to better serve the web-based application and effectively manage and report the website. There are various traditional methods like LRU, LFU etc. for web proxy caching. The web log files may be preprocessed to find meaningful patterns and improve the hit ratio of the web server. The proposed study used the web proxy data of directorate of distance education, Dibrugarh University for data mining and analyzing the performance of the proposed algorithm on the web proxy server.

Keywords: *Web Mining, Web Proxy Cache, Bayes Theorem, Algorithm, Pattern Extraction*

1. INTRODUCTION

With the explosive growth of data in Internet, it has become important to find the useful information from these data. Web Mining may be described as the extraction of useful information from the web documents and services by using data mining tools and techniques [1]. The paper has two major components. First Component is to mine the Directorate of Distance Education (DDE), Dibrugarh University website and find out some interesting results. The second component is to develop an algorithm for improved web proxy cache management.

The Directorate of Distance Education (DDE) was established in Dibrugarh University in the year 2000 to cater to the growing needs of Higher Education in this eastern-most part of the country. The Directorate of Distance Education is offering at present MA Programmes in Assamese, Economics, English, Political Science, Sociology, M.A /M.Sc. in Mathematics, M.Com; and One-Year Post Graduate Diploma Programmes in Marketing Management (PGDMM) and Journalism and Mass Communication (PGDJMC). Besides these, BBA, BCA and PGDCA programmes have been introduced from 2009-2010 sessions. At present, there are more than Sixteen Thousand learners enrolled in the Directorate.

For mining DDE, Dibrugarh University website, the authors used web log expert software. For web proxy cache management, the authors developed an algorithm based on naïve bayes classifiers and bayes' theorem.

This paper is organized as follows: Section 2 discusses the methodology. In Section 3, the directorate of distance education, Dibrugarh University was mined and interesting facts are presented. In Section 4, the authors proposed a better performance algorithm. Section 5 evaluates the proposed algorithm. Section 6 concludes this paper.

2. METHODOLOGY

2.1 Web Mining

Web is a collection of sea of information and documents. The Web is very dynamic and flexible in nature. The World Wide Web grows in respect of traffic, size, data and complexity. It is difficult to extract the relevant and interesting information from the huge amount of web data. In this regards, data mining plays a major role. Web mining consists of three types of processes – 1.Web Content Mining, 2. Web Structure Mining, 3.Web Usage Mining (as shown in Fig. 1). The Web Content Mining deals with the contents of the web pages and structure mining concerns with the structure of the website. Web Usage Mining is used for the information related to the users' characteristics and their usage of the websites. The semi structured format of the web usage mining needs pre-processing and parsing before gaining the hidden information. The term web usage mining was coined by Cooley in 1977 [2].

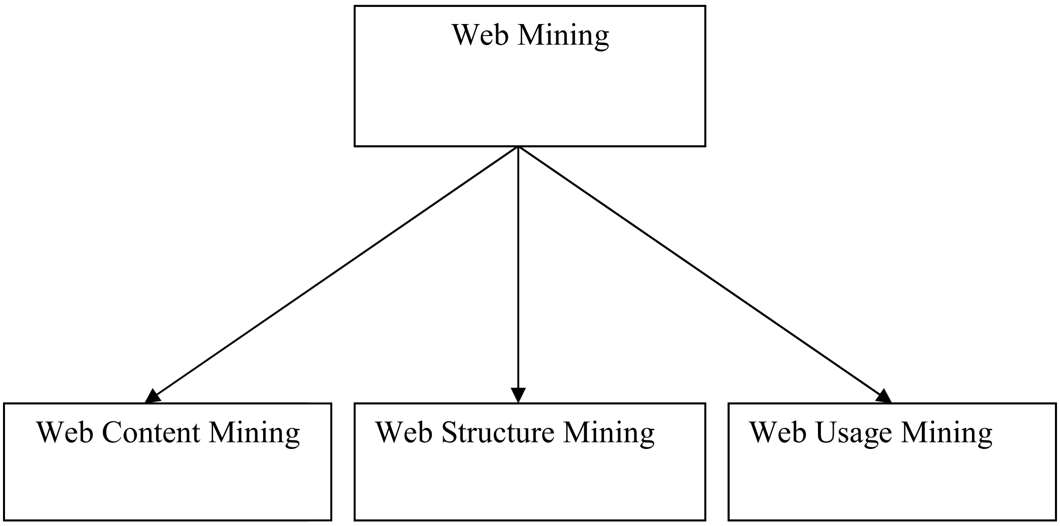

Fig. 1. Web Mining And Its Different Categories

The web cache algorithms play a crucial role in web mining [3, 4]. Many web proxy policies try to gather information on the factors that can influence the decision of caching [5]. The important decision is that which web object needs to be cached and which object needs to be replaced to make use of the available cache size, hit rates, reduce network traffic and minimize the loads the web server. [4, 6] In web proxy server, the logs are the records of the user activities and may used for prior knowledge of the future accesses. The web environment changes very rapidly and continuously. Most of the recent studies used artificial neural network as caching technique [6, 7]. But it has the drawback of extra computational overheads as the ANN training set may consume considerable amount of time. The Bayesian network is used in many fields such as forecasting, military applications, statistics, control and modeling [8]. Naïve Bayes classifier is applied in many fields. So, Bayesian approach may be utilized to produce better solution for web proxy caching.

2.2 Weblog Expert

WebLog Expert comes in two flavors – “Lite” and “Full” Editions. It is a purchased application with 30 days trial and works with web servers such as Apache and IIS. WebLog Expert generates various statistical information and reports based on the logs files. It provides the number of visitors, unique IP, bandwidth, spider application, keyword searched, hit ratio etc. [9].

2.3 Bayes' Theorem

The concept of conditional probability is introduced in Elementary Statistics. The authors observed that the conditional probability of an event is a probability obtained with the additional information that some other event has already occurred. The authors used P(B|A) to denoted the conditional probability of event B occurring, given that event A has already occurred. The following formula was provided for finding P(B|A):

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

The conditional probability of B given A can be found by assuming that event A has occurred and, working under that assumption, calculating the probability that event B will occur. The authors extends the discussion of conditional probability to include applications of Bayes' theorem (or Bayes' rule), which is used for revising a probability value based on additional information that is later obtained. One key to understanding the essence of Bayes' theorem is to recognize that we are dealing with sequential events, whereby new additional information is obtained for a subsequent event, and that new information is used to revise the probability of the initial event. In this context, the terms prior probability and posterior probability are commonly used.

Definitions A *prior probability* is an initial probability value originally obtained before any additional information is obtained.

A *posterior probability* is a probability value that has been revised by using additional information that is later obtained.

The probability of event A, given that event B has subsequently occurred, is

$$P(A|B) = \frac{P(A)\cdot P(B|A)}{[P(A)\cdot P(B|A)]+[P(\overline{A})\cdot P(B|\overline{A})]}$$

2.4 Naïve Bayes Classifier

Naïve Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness and diameter features.

For some types of probability models, naïve Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naïve Bayes models uses the method of maximum likelihood; in other words, one can work with the naïve Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers.[10] Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.

An advantage of naïve Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification.

2.5 Web Proxy Cache

Caching is an effective performance enhancing technique that has been used in computer systems for decades. However, proxy caching differs substantially from the traditional ones used in processors. Two main features of the World Wide Web applications that differ from the traditional caching are (i) non-uniformity of the object sizes and (ii) non-uniformity of the cost of cache misses (as opposed to the traditional caching where all cache blocks have the same size and require the same amount of time to be retrieved in case of cache misses). The traditional metrics for measuring caching efficiency has been hit-ratio (HR), which is defined as the number of requests satisfied by the cache divided by the total number of requests. Obviously, HR is not an appropriate metric to measure performance of proxy caches, because of the non-uniformity of object sizes and non-uniformity cost of misses. Byte Hit Ratio (BHR) is defined as the number of bytes found in the cache divided by the total number of bytes requested within the observation period. The BHR takes into consideration the non-uniformity of the object sizes, but it fails to consider the non-uniform cost of misses. Another phenomenon is that users' interests overlap in time, which indicates that part of the locality observed by the proxy comes from the fact that the proxy sees a merged stream of accesses from many independent users, who share a certain amount of common interests. Figure 2 depicts the basics web proxy cache environment.

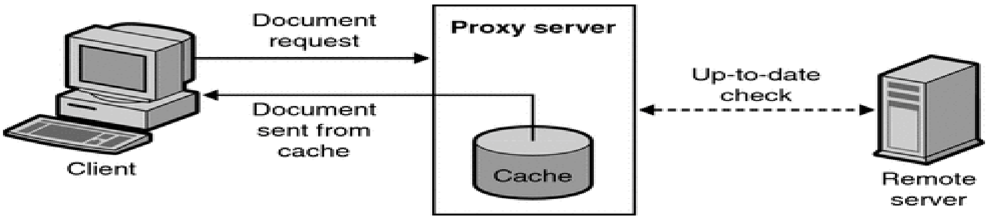

Fig. 2. Web Proxy Cache

2.6 Squid Native Access.Log Format In Detail

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

1. **TIME** A Unix timestamp as UTC seconds with a millisecond resolution. This is the time when Squid started to log the transaction, which normally happens at the end of a transaction lifecycle, after the entire request was received from and the entire response was sent to the HTTP client.

2. **DURATION** The elapsed time considers how many milliseconds the transaction busied the cache. It differs in interpretation between TCP and UDP:
For HTTP this is basically the time from having received the request to when Squid finishes sending the last byte of the response.
For ICP, this is the time between scheduling a reply and actually sending it.

3. **CLIENT ADDRESS** The IP address of the requesting instance, the client IP address.

4. **RESULT CODES** This column is made up of two entries separated by a slash. This column encodes the transaction result:

The cache result of the request contains information on the kind of request, how it was satisfied, or in what way it failed. Several codes from older versions are no longer available, were renamed, or split. Especially the ERR_ codes do not seem to appear in the log file any more. The status part contains the HTTP result codes with some Squid specific extensions. Squid uses a subset of the RFC defined error codes for HTTP.

5. **BYTES** The size is the amount of data delivered to the client.

6. **REQUEST METHOD** The request method to obtain an object. If one turned off log_icp_queries in configuration, one will not see (and thus unable to analyze) ICP exchanges.

7.**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.

8.**USER** The eighth column may contain the user identity for the requesting client. This may be sourced from one of HTTP authentication, an external ACL helper, TLS authentication, or IDENT lookup (RFC 931) - checked in that order with the first to present information displayed. If no user identity is available a "-" will be logged.

9.**HIERARCHY CODE** The hierarchy information consists of three items:

Any hierarchy tag may be prefixed with TIMEOUT_, if the timeout occurs waiting for all ICP replies to return from the neighbours. The timeout is either dynamic, if the icp_query_timeout was not set, or the time configured there has run up. A code that explains how the request was handled, e.g. by forwarding it to a peer, or going straight to the source.

10.**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. For requests sent to a neighbor cache, this is the neighbor's hostname.

11.**TYPE** The content type of the object as seen in the HTTP reply header.

3. DDE WEB MINING USING WEBLOG EXPERT

Using WebLog Expert, the authors analyzed the statistical reports for the period 28.02.2015 to 15.3.2015 and found that index.htm is the most popular page which is obvious. The page.php and newsdetails.php are the next popular pages. The figure 3 depicts the most popular pages. From the figure 4 given below, one may say that the Android is the most popular OS used for accessing DDE, Dibrugarh University website. As the DDE, Dibrugarh University does not have an android web application till date; it is highly desirable to have one according to the statistics. The figure 5 shows the top keywords through which the website is searched. The keyword ‘ddedu’ tops the chart. The total hits for the website during the period is 423,320 and cache request is 31,667. The total visitors are 17,704 and on the other hand, Unique IPs are 8,874. The total bandwidth is 9.83 GB and average bandwidth per day is 629.10 MB.

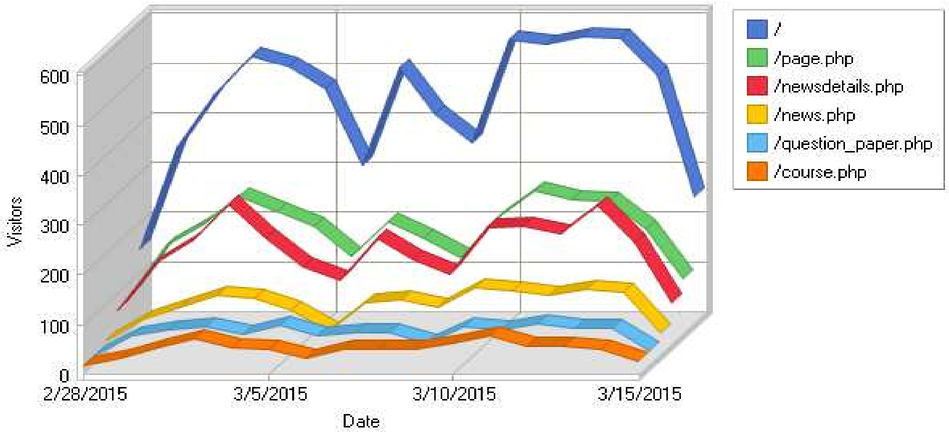

Fig. 3. Most Popular Pages In The Time Interval

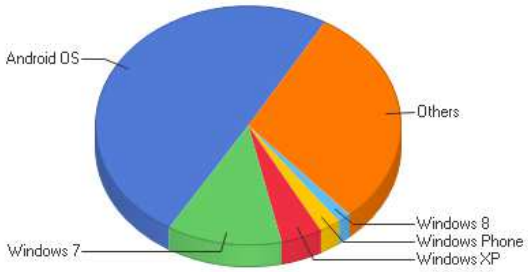

Fig. 4. Most Popular Operating System

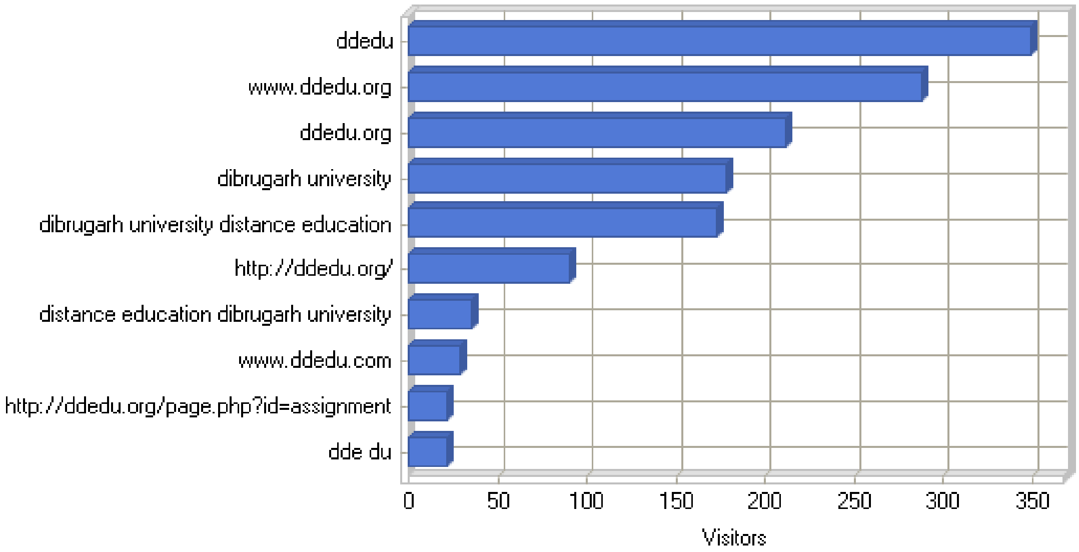

Fig. 5. Top Search Phrases

4. PROPOSED ALGORITHM FOR IMPROVEMENT OF WEB PROXY CACHE MANAGEMENT

ALGORITHM FOR WEB PROXY CACHE MANAGEMENT

```
Begin

        Data_cleaning();

        Classify the Web Objects using Naïve
Bayes Classification

         Train the datasets and move the objects
according the classifiers and assign priority
weightage of the classes.

          Find the prior probabilities of the web
objects P(Wi) where $W_1$, $W_2$, …$W_n$ are the web
objects of the Web Server from the different
classes.

          Find the posterior probabilities  Post($W_i$)
using Bayes' Thoerem

        Post($W_i$) = apply_bayes (common features)

        For each web object $W_i$

         Mark the Web object $W_i$ according to the
Post($W_i$) priority wise in descending order for
different classes
Insert the Web object $W_i$ to the priority queue
Delete the Web object $W_i$ from the priority queue
based on class priority weightage and the posterior
probabilities.
       If cache miss occurs
        Find the web object from the original server
      End
     Refresh the cache buffer after the pre-
determined time
End

       Data_cleaning()

 Begin
        Process the status codes, scripting codes
generated automatically, picture file codes, Style
Sheet file extension, Size of the web object present
in the logs
       The threshold value for each of the web
object is decided.

        The selected files and objects are cleaned
from the log file.
End
```

5. EVALUATION

The authors evaluated the algorithm using simulator called Web Traffic Generator. The comparison was made using both the traditional as well as the proposed algorithm method. The performance of the web proxy caching had certainly improved if one analyzed Hit Ratio and Byte Hit Ratio.

	Web Proxy Server using traditional Method	Web Proxy Server using the proposed algorithm
Total requests	110694	110694
Cacheable requests	52436	52436
Unique requests	48013	48013
Max HR (%)	15.67	21.88
Max BHR (%)	19.34	23.49

6. CONCLUSION

The study had analyzed the Directorate of Distance Education, Dibrugarh University website and found some interesting information out of it. The authors also proposed an algorithm for improved web proxy caching. The traditional caching policies are still used in most web proxy servers. The most common web proxy caching are Least-Frequently-Used (LFU), Least-Recently-Used (LRU), SIZE policy etc. There are some research works done in this area and Weleed et. Al. [11] proposed an intelligent naïve bayes-based approaches for web proxy caching. The authors also proposed another algorithm in this regards and this is high time to go for the change for better performance.

7. ACKNOWLEDGMENT

The authors expressed their gratefulness to Prof. Alak Kr. Buragohain, Vice-Chancellor, Dibrugarh University for his inspiring words. The authors also acknowledged Prof. Nirode Boruah, Director i/c, DDE, Dibrugarh University for allowing using the log data of the website www.ddedu.org. The authors also expressed their gratitude to Mr. Abhijit Adhyapak, Asstt. System Analyst, Dibrugarh University in performing the experiments.

REFERENCES

[1] O. Etzioni, The World Wide Web iQuagmine or gold mine, Communications of the ACM, 39 (II), 65-68, 1996.

[2] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, “Grouping Web page references into transactions for mining World Wide Web browsing patterns”, 1997.

[3] H.T. Chen, Pre-fetching and Re-fetching in Web Caching Systems: Algorithms and Simulation, TRENT UNIVERSITY, Peterborough, Ontario, Canada, 2008.

[4] T. Chen, Obtaining the optimal cache document replacement policy for the caching system of an EC website, Eur. J. Oper. Res. 181 (2007) 828-841.

[5] J. Cobb., H. ElAarag, Web proxy cache replacement scheme based on back-propagation neural network, J. Syst. Software, 81(2008) 1539-1558.

[6] Waleed Ali, Siti Mariyam Shamsuddin, Neuro-fuzzy system in partitioned client-side web cache, Expert Sys. Appl. 38(2011) 14715-14725.

[7] Waleed Ali, Siti Mariyam Shamsuddin, Intelligent client-side web caching scheme based on least recently used algorithm and neuro-fuzzy system, Advances in Neural Networks ISNN 2009, Publishing, 2009, pp. 70-79.

[8] C.-G. Bai, Bayesian network based software reliability prediction with an operational profile, J. Syst. Sotware, 77 (2005) 103-112.

[9] V. Jayakumar and Dr. K. Alagarsamy, “Analyzing Server Log File Using Web Log Expert In Web Data Mining”, International Journal of Science, Environment and Technology, Vol. 2, No 5, 2013, 1008 –1016.

[10] Zhang, Harry. The Optimality of Naive Bayes. FLAIRS2004 conference.

[11] Waleed Ali, Siti Mariyam Shamsuddin, Abdul Samad Ismail, “Intelligent Naïve Bayes-based approaches for Web Proxy Caching, Knowledge-Based Systems 31(2012) 162-175.

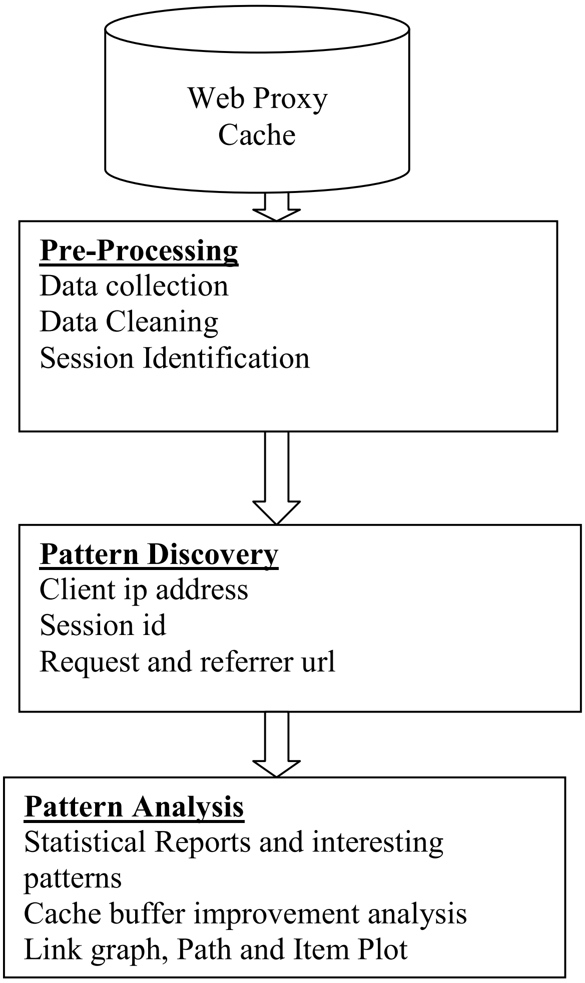

Fig. 6. Web Proxy Cache Mechanism Used For Pattern Discovery