Bachelor's Thesis(UAS)

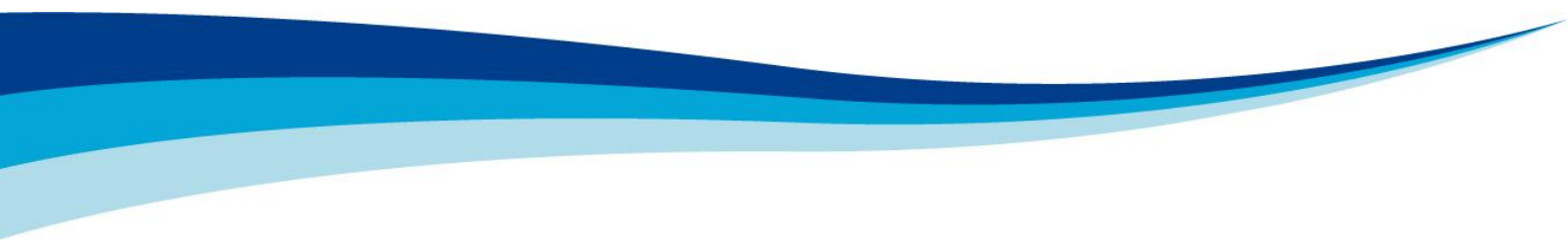
Bachelor of Engineering

Information Technology

2012

Ezekiel Ufwinki

Web Log Pre-processing



TURUN AMMATTIKORKEAKOULU
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR´S THESIS | ABSTRACT
TURKU UNIVERSITY OF APPLIED SCIENCES

Degree programme | Information Technology

Completion of the thesis| 26

Instructor(s) Patric Granholm

Author(s)Ezekiel Ufwinki

WEB LOG PRE-PROCESSING

Over the past decade, with the rapid growth in Internet, especially Web2.0 era and BS application times, the arrival of blogs, virtual communities, online office, e-commerce, e-government, B2B and C2C and other emerging Web applications, the Web has become one of the core elements of human life and work. How can we enhance the value of the Web site, allowing users a better experience, and quickly find the information we need to find the user's needs? How can we improve the competitiveness of e-commerce applications and to survive in the fierce war of the Internet? These issues require answers we can find in the vast amounts of Web data. Thus, the combination of data mining technology and Internet applications constitute a very active and very important a field of study, in other words, Web mining.

Having a similar structure and content of the access log file on each Web server, Web logs automatically become an important data source for Web mining and its mining has a universal and practical significance. However, the large amount of web log data, containing a lot of noise, not suitable for Web mining, must first be pre-treated.The workload of data pre-processing accounts for more than 50% of the total web mining workload. This thesis introduces the Web log, the log pre-processing methods, and seeks the maximum forward path and frequent traversal path algorithm based on the use of http://shopping.yahoo.com/.

KEYWORDS: data mining, data pre-processing, Web logs, Web mining

Content

LIST OF ABBREVIATIONS (OR) SYMBOLS

HTML	HyperText Markup Language
CLF	Common Log Format
ECLF	Extended Common Log Format
ExLF	Extended Log File Format
HTTP	Hypertext Transfer Protocol
MFP	Maximum Forward Path

1. INTRODUCTION

In this internet era web sites on the internet are a useful source of information in day-to-day activities. So there is a rapid development of the World Wide Web in its volume of traffic and the size and complexity of web sites. As per August 2010, according to the Web Server survey by Netcraft. there are 213,458,815 active sites. Web mining is the application of data mining, artificial intelligence, chart technology and so on to the web data and traces user visiting behaviors and extracts their interests using patterns. Because of its direct application in e-commerce, Web analytics, e-learning, information retrieval and so on, web mining has become one of the important areas in computer and information science.

During the time of Web mining, Web applications are not the same, but each Web server has a structure similar to the access log file, so its excavation has a general and realistic significance. In this thesis Web logs, without special instructions, refer to the Web server side of the access log.

Of course, to carry out excavation for a specific Web application, the best and most accurate method is to build Web applications. Web mining needs to take into account the useful information through the Web application records or the custom format log data. However, the Common Log Format (CLF) method does not have universal significance, so this is not to be discussed in this thesis.

By mining Web server log files, we can identify the paths that user groups used to access the web page This is known as user clustering analysis and it helps optimize the access path and thus improve site topology. In addition, web server log files help us identify the content that user groups are accessing. This knowledge enables web developers to provide personalized services. Personalized services means that we offer web content and links tailored to the interests of the users. Furthermore, analysis and research of the user behaviours can lead to developing a marketing strategy for potential users which can lead to greater competitive advantage.

Therefore, the Web log mining technology has important significance, on the following aspects:

a .Web Personalization

The process of providing information that is related to a user's current page is known as web personalization. This information is usually displayed on the current page in the form of web page links. The idea behind web personalization is that the web page currently being browsed by a user indicates his/her interest in that topic and it is likely that the user would be interested in similar information. For example, in case of e-commerce the related information could be about other similar products to those that the user is viewing or about products that other users who bought or viewed this product also bought.This example would also work for a research or target-oriented web browsing.

The key information that is required for suggesting these similar web pages comes from the knowledge of other users who have also visited the current page as well as other pages before and after this current page. In addition to other users' browsing information, web personalization can also take advantage of the web page content, the structure of the web page or the user's profile information. All these help in creating a focused and personalized web browsing experience for the user.

b.. System improvements

The original design purpose of web server logs was to provide statistics for Web site management and system administrators. Analyzing the log helps to better study the Web caching, network transmission,load balancing,data distribution strategy,leading to a conclusion for the Web system performance improvements.Web traffic behavior analysis is used to achieve a balance of access, reduce congestion and optimize the transmission. In addition, the analysis of unusual large-scale traffic and frequent access error can prevent web site intrusion, deception, and removal of invalid links.

c. Website structure design

Web log mining for Web site designers provides detailed user feedback to help them to adjust the topology of the structure and content of the Web site, according to the actual user's browsing, and optimize the Web site, in order to better serve users.

d. helping business decision-making

Concerning e-commerce sites, analyzing the log, the user buying trends, studies of user psychology, business decisions, or through the analysis of the source URL, adjusting the web input, effectively increases site traffic.

e. Search Engine Optimization

Search engines better index websites by analyzing the behavior of the Web Crawler Web log, a site of structural adjustment.

Data mining requires data pre-processing because the data in the real world is mostly incomplete, noisy and inconsistent and in a variety of data formats. For data mining algorithms, incorrect input data may lead to wrong or inaccurate mining results; at the same time, the data mining algorithms are usually dealing with a fixed-format data and as the data exist in reality in a wide range of formats, we need to process these data before we can use these data into the data mining algorithms. Data mining algorithms may be only part of the data in the database mining, and because of this, we need to extract useful data. To fix the real world data which are incomplete and inconsistent, we need to remove noisy data, and convert the data coming from multiple sources into a consistent the format. Data preprocessing accounts for 50% of the entire data mining process and the results of data preprocessing are the input of data mining algorithms which directly affect the quality of the mining At present, researchers have proposed many effective data pre-processing technologies. Common data cleaning (Data Cleaning) removes the noise from the data and corrects data inconsistencies; Data Integration (Data Integration) combines multiple data sources into a consistent data storage; data transformation (Data Transformation) and data protocol (Data Reduction) can be gathered to remove redundant features or and the clustering method compresses data. Before data mining, data preprocessing techniques can greatly improve the quality of the data mining model and reduce the time required in the actual digging and disk space. In other words, data pre-processing can improve the quality of the data which helps to improve the accuracy and performance of the subsequent mining process. Quality decision-making must trust the quality of data, therefore, data preprocessing is an important step for the knowledge discovery process. Detection of abnormal data, as soon as possible to adjust the data, and the

Statute of the data will yield good returns in the data mining process.

The purpose of the Web log mining data preprocessing is to remove useless data from the Web log mining process, and place Web log data into a recognizable form for the mining algorithm.

Web Log Preprocessing includes the following steps: data cleaning (Data Cleaning), user identification (Users Identification), session identification (Session Identification), the path completion (Path Completion) and transaction identifier (Transaction Identification).

This thesis introduces the general approach of the major steps of the Extended Common Log Format (ECLF) log preprocessing method in the second chapter. ECLF is used for calculating the maximum forward path and frequent access path algorithm. Chapter 3 will give an example of preprocessing of a web log, by first setting up a data warehouse, and logging into the database, and then cleaning up the analytical work. Chapter 3 also introduces user clicks on the event model, and proposes an algorithm for this model, that is, the maximum forward path and frequent access path.

2. WEB LOG PREPROCESSING

Web Log Preprocessing includes the following steps:

- data cleaning (Data Cleaning),
- user identification (Users Identification),
- session identification (Session Identification),
- the path completion (Path Completion) and
- the transaction identifier (Transaction Identification).
  Transaction identifier means that we are given a maximum access path before the path (maximal forward references) of the algorithm. Demand frequent access path (Large reference sequences) is the parameter frequently used in web mining and it establishes the maximal forward path (maximal forward references). On the basis of it and after introducing the general preprocessing process, we will give it the general algorithm

2.1 Web Log Format

The most commonly used Web server software uses one of the three kinds of open log file format to record the log file. These three kinds of file formats are:

a. CLF (Common Log Format),
b. ECLF (Extended Common Log Format), and
c. the W3C Extended Log File ExLF (Extended Log File Format).

The Extended Common Log file Format is displayed in Table 1. If the Web server domain data is unavailable, then the Web server will be the successful tenderer in this airspace, "-".

Table 1. Extended Common Log Format ECLF

Remote host domain	User submits the request host name, the general record of the IP address.
rfc931domain	System identifies the user remote login name from the multi-user system; it almost always contains a "-" symbol.
Authorized user domain	It saves the http user authentication user name.
Date domain	It requests the date and time.
Request domain	HTTP requests from clients for this request to establish the first connection. If the requested file exists, this field will determine the URL of the requested files, and access to this file.
Status domain	This is the status code: the status code of this file is requested successfully.
Bytes domain	This is statistical data comprising of domain byte requests, and does not include the HTTP header information.
Referrer domain	We can go to this page by clicking the link to the URL of the page. If this link does not exist, it is saved as "-". The domain data is actually extracted from the HTTP header Referrer domain; this contains the Referrer HTTP along with the page request sent with.
User Agent domain	This refers to requesting the name and version of browser from the HTTP header user agent field.

The Status domain has a total of five categories of status codes which are used by system administrators and developers to provide information:

1 ,. 100 indicates continue and 101indicates protocol conversion;
2 . 200indicates that the operation was successful.
3 196: indicates that the requested resource exists in another URL.
4 404 indicates that there is an error. The most common code is 404: File not found.
5 500 indicates that the request cannot be carried out because of network problems or your response. 500: Internal Server Error.

The reference domain and the user agent field is added ECLF which is relative to the CLF.

The ExLF format is not used and is, therefore, not described here.

The Request (Request) domain contains the request method and the requested resource URL, has the OPTIONS request: GET, HEAD, POST, PUT, DELETE, TRACE, and CONNECT. Here we are concerned with the GET method which retrieves the URL to identify the resources.

The Agent domain can be used to identify the browser. The Web program can optimize the different browsers based on this domain, and can also be used to identify some of the log data through the domain, such as YahooSeeker/1.2 (compatible with Mozilla 4.0; MSIE 5.5; yahooseeker at yahoo-inc dot com; http://help.yahoo.com/help/us/shop/merchant/).

2.2 Data cleaning

The Web log contains information for each http request, but not every data after the mining is meaningful. For instance, a user requests a page and while browsing the page can also download pictures, video, CSS files, JS files. Data cleaning gets rid of these requests, reducing the amount of data.

Data cleaning can be carried out under the following three aspects:

- URL: website. As the HTML file is related with the user session, the suffix for gif, jpg, js files can be filtered from the log. For some special sites, such as photo site, we can reset the relevant information. But for some dynamic websites, all content is dynamically generated and filtering rules must be adjusted according to the procedure.

- The requested action: We can only retain the GET action.
- Return status: requesting successful track record can only keep a record of 404,501, etc., and return the error code to be removed. The error log on the website maintenance and security analysis is very important, if the system administrator should analyze these records.

Request IP: to get rid of the access from the network robot, we can create a robot IP filter list.

Experiments on 550.8M (3,078,210 rows) log files to clean up the results obtained 44.4M (265,016 rows) result in reductions in data volume.

2.3 User Identification

The user refers to an individual accessing one or more servers through a browser. Due to the presence of caching, firewall and proxy server, the only reality is that it is very difficult to identify a user. A Log can distinguish the user's user IP, browsing device operating system identification and session cookies.

Because multiple users may be accessed through a proxy, a single IP corresponds to multiple users, and it is difficult to distinguish between users via IP.

In browsers and operating systems with IP there are some difficulties as the user ID of the user's operating system and browser is more concentrated so a large number of users using the same IP cannot be distinguished.

Using session cookies to assign each user a unique identity relates to user privacy issues, and the user may simply not support cookies, or the user will delete or modify the cookies, so session cookies are not trustworthy. Cookies can be retained on the server side in order to accurately identify the user session information, including the session ID, user name of a registered user visiting the page. Some Web servers such as Apache record cookie data with the help of a number of modules. If a web server does not record cookie data, it then needs the support of a web application, otherwise versatility is not high.

This thesis discusses the general log preprocessing only due to data limitations, which are imposed by the rules of identifying users. For example, if a user's IP is different, then this user is a different user; if the user IP is the same but is a different Agent, then, the user is also seen as different users. If the above two are the same, the user cannot access the page history the page of history.

2.4 Session Identification

The session identification is the user's access records in a single session. A timeout mechanism tends to be used to divide the session. If the difference between the two-page request time exceeds certain limits, then the user needs to start a new session.
Many web applications use 30 minutes as the default timeout (as is the PHP default session timeout value).

2.5 Path Added

Path added or path completion is the process of adding the page accesses that are not in the weblog but that have actually occurred. In one session, if there is a request from the previous page, then, the previous page is added as the source of this request. If a user uses a number of pages to reach to the final page, then the last page before the final page becomes the source page and it is referred to as the Referrer domain. For example, if the users goes to the bbc news webite and from there pick news about Africa and from there he chooses to read a sports story. Then the news about Africa is the referrer domain. 2.6 2.6 Transaction Identification

Each user session can be seen as composed of multiple transactions, a transaction is a group of a certain semantic history data.

A major transaction identification method is to find the session prior to the path (maximal forward references, MFP), each MFP is a transaction. MFP is defined as a group prior to the browsed page. The request page is not the visited page, “back” refers to the accessed page in the history of the user session prior to the visit. There will be new pages added to the traversal path, while “back” does not extend to the user's access records.

2.7 Maximum Forward Path

Assuming that there has been a session identification, a group of session file, the session traversing the path {the X1, ... Xm}, MF, = {Y1, ... Yj} potential MFP, is initially empty, MFSet as {MF1, ... MFn}, then the path for the session maximum before collection, is initially empty. Flag signs indicate the current traversal direction is forward or backward. The following is an evaluation of the MFP algorithm:

for each {X1, ... Xm} Xi

1) if Xi is not MF , then Xi as the last element of the MF to join,

Otherwise there is Xi = Yk, 1 <= k <j-1,

(a) If the flag is forward, the current MF added MFSet as an MFP, and then delete a page from MF {Yk +1, .., Yj-1}, and set the flag for the back, into the next cycle

(b) if the flag is back, this time the MF is not the MFP directly to delete {Yk +1, .., Yj-1} into the next round of the cycle.

If the loop is the last page of the user session, the flag is still forward, then the {Y1,.. Yj-1} is one the MFP and it is addedto MFSet.

2.8 Frequent Access Path

Frequent traversal path is a continuous page sequence of the MFP support for more than a certain threshold. The number of user sessions containing a frequent traversal path is called the support. The definition of frequent traversal path length contains the number of pages. For FPk Hutchison length = k frequent traversal path collection, a collection of one of the most frequent traversal path of M FPk, M = {Pk, 1,.., Pk, M} of elements of support from large to small order.

2.9 Web Log Preprocessing Major Challenges

- The two major challenges involved in Web usage mining are preprocessing the raw data to provide an accurate picture of how a site is being used, and filtering the results of the various data mining algorithms in order to present only the rules and patterns that are potentially interesting.
- Log Credibility

The user can modify the HTTP header sent, which is seen by the user as a browser program, thus affecting the referrer domain and agent domain, so the only way to get real customer behavior is to reach the user through the web application track record of work. Log files will not be credible ifmost users still use the browser's default configuration. A fake user can filter out data cleaning and interference information does not leave on our analysis of the result too a big impact.

- Dynamic Web sites

In today's web applications, static web sites has been very scarce; dynamic sites, may contain a long string of requests, some for the page content, while others are some sessions. The id request parameter may represent a different URL page, such as / doc.htm? user = me function = good / doc.htm? the user = you & function = good point to the same page. The function parameters determine the page content, user parameter only saves the user's id; also there may be image requests by program indirectly fulfilled such as / getPicture.jsp? picid = 95 536 This is actually a request for an image and this time Web preprocessing is necessary to take into account these factors.

- Other issues

Log data are huge, therefore, we need to consider the consumption of CPU and IO in the preprocessing process.

3. PRE-PROCESSING SYSTEM IMPLEMENTATION

This chapter proposes a Web log (ECLF format) preprocessing process; this implementation includes the establishment of a log data warehouse, the log data cleaning, session identification, analysis function of the forward path and frequent access path. One of the biggest forward path and frequent access path algorithm is based on the user clicks on the event model.

3.1 User Clicks on Event Model

Information from a web log records can be requested URL and Referrer URL (reference URL), each one logging as user clicks on the event. The referrer domain and the request domain are equally important from the point of view of web mining. The referrer domain above the current request, such as a web page A, clicks on the home page link to the homepage A to enter the page A and clicks on the page sidebar or through a search engine on the homepage A. So by clicking on different links on webpage A, a user displays a click behavior. The referrer domain and the current page request determine the user clicks on the event so that we find the basis of the forward path and frequent access path. Below we will use the requests, the referrer format to determine a Web click event, such as a.html b.html a user clicking on the user request a.html, and a reference page for the b.html.

In one session, the referrer field of an event A is equal to the request of another event B domain (event A occurs after event B), meaning that the page references the event A on the event B page. In general, the most recent event B is after the event A, and event B is the result of the event A.

The proposed model and the reasons why it is proposed are:

The path and frequent access path are based on the maximum forward path before the click event model derived on the Web mining deeper meaning, ignoring the context of the user clicks on the event. A browser such as the following sequence: (a.html, -), (b.html, a.html), (c.html, b.html), (a.html, c.html), (d.html, a html) in the implementation will be two of the largest before the path a.html-> b.html-> c.html, a.html> d.html, .Users second request a.html was back, click on the event model, it will only get a maximum forward path (a.html, -) -> (b.html, a.html) -> (c.html, b.html) -> (a.html, c.html) -> (d.html, a.html), which is good when this path is frequently used as we will be able to join in c.html ,d.html link the user without going back to a.html nonstop d.html, but

also to reduce the burden on the server. We using this method to get out of the two paths and in the later analysis we will see such a relationship. Sometimes the program may ask Who? In this case, clicks on the event model of the analysis target is the user behavior as much as possible cascading into a linked continuous flow, while with the former approach it is easy to lose the relationship between user behaviors.

Second, the referrer domain gives information, and simplifies the process. In the previous implementation, it is necessary to know whether a page from another page needs to refer to the site topology. First of all, with this structure is difficult to obtain the structure of the site which may continue to adjust. Especially for dynamic sites, this structure is almost impossible to deal with some history log so the topology may be totally useless. Modern browsers (except for special configuration and disguised as a browser program) will normally be included in each request referrer domain, and web log records this domain, so the use of this domain can simply be the relationship between the page, if the referrer is the request for a.html b.html you can believe there is a link to a.html b.html in.

Effectively ECLF filters the same IP with an agent of multiple users. In the concept of context, we can further distinguish between multiple users that may exist so some requests on the topology institutions are interrelated, in fact, they are requested by different users through the Referrer to a certain extent, to avoid seeking the most forward path.For convenience, the system analysis and processing external URL, the referrer field is empty or the referrer domain for the external URL of the site is seen as the same as the external referrer. As a result of clicking on the event model, as mentioned above, the maximum forward path algorithm and the frequent traversal path algorithm and the previous mentioned are different.

3.2 Creating a Database

The general data cleaning process is to log data, such as a data warehouse, into a log of pre-treatment system whose goal is to get the data warehouse cleaned for mining use.

LOG_MAIN is the main table which stores the general information of the logging, web log records corresponding to a data in the database table.

The design purposes of this database are to save in the log all the data, provide some data redundancy, to facilitate later analysis in the loading process

as much as possible and save some auxiliary information

3.3 Data Sheet Features

Save the remote host domain log REMOTEHOST IP_ADDRESS is to save the user IP in the table which is only IS_ROBOT identifies the user is not a web robot, the table used to store the list of network robot, combined with the data cleaning IS_ROBOT field. IP_1, IP_2, IP_3, IP_4 store four IPv4 domains when used to ECLF filter as an IP segment.

LOCAL_URL logs the requested URL, LOCAL_PATH saves the URL relative path, that is, to get rid of the URL of the domain. IS_CONTENTPAGE , IS_VALID are two fields that identify the record content page, not digging a valid page for filtering. The IS_HOMEPAGE identity is home.

AGENT: It saves the agent domain information in the log, AGENT_FIELD storage agent domain, IS_ROBOT identifies that the agent is not a web robot and are used for filtering. BROWSER, BROWSER_FAMILY, OS, OS_FAMILY information is stored on the browser and operating system for future mining.

USERS: For user identification, the same IP address and the same Agent are identified as the same user. REMOTEHOST_ID, AGENT_ID points to REMOTEHOST table and AGENT table. IS_VALID mark is a valid user for data cleaning.

LOG_MAIN: It logs information. REFER_FROM_LOCAL says that the referrer domain is not a local URL, if it is REFER_URL_ID to point LOCAL_URL of a record, REQUEST_URL_ID, always points to LOCAL_URL a record.

ACCESS_LOG: It corresponds to user clicks on the event. HITS on behalf of the Hits, FROM_URL_ID on behalf of referrer page, point to the LOCAL_URL of a record, if it is an external URL or does not exist with 0. REQUEST_URL_ID is used to point LOCAL_URL of a record, on behalf of the requested URL. FROM_URL_ID, REQUEST_URL_ID in this table only.

MF: This is the path to save the maximum forward path of all users. PAGES path length, DIRECT_SUPPORT SUPPORT saves support. The MF field saves the path information in this table only.

SESSION_MAIN: It records session information, for example, USER_ID of the user of the session.

SESSION_LOG: It saves the session hits record, mainly used for the

middle of saving data analysis and calculation, the SESSION_ID points to a session, LOG_ID points to a logging (LOG_MAIN.LOG_ID), a record of ACCESS_LOG_ID points to ACCESS_LOG.

SESSION_MF: It saves the session path information. MF_ID points to the record in an MF.

3.4 Data loading process

Data loading for each step is as follows:

(1) establishing the secondary data, stored in the Agent table search engines and the robot's IP.

(2) reading and analyzing log files; each log record is written to the database.

Some simple filters, such as that the request domain is not the resource of the site, getting rid of the non-GET access, removing the return code is 4 ** 5 ** access, t removing f access to the IP domain of a few common search engines, getting rid of the access of common media file extension. Where the data cleaning can be effective is in reducing the amount of data written to the database. Of course, if we do data cleaning for a variety of mining data warehouse, we can get rid of the clean-up steps, or add a field in the data table LOG_MAIN, logo for deletion, in fact, saves all data.

Each logging domain of the URL of the Remote Host, agent, requests the first query in the database relevant records if there is no record of the new record, then it is inserted. At the same time for new Remote Host and agent combinations in the USERS table, a new record is inserted. First when processing the URL to the URL of relative path, we also deal with issues such as / / dd / a.html and / dd / a.html is actually the same page with other issues.

3. Data cleaning

Detailed cleanup is removing a more comprehensive list of network robot IP access from the network robot by analyzing the Agent domain. Robot and program access, and the analysis of the saved request URL help to clear out meaningless pages, documents.. The first bid step is to increase the LOG_MAIN field identifiers, rather than directly delete them.

4. Subscriber Identity

To uniquely identify a user, the IP address and the Agent field, in the finishing of the above data analysis, have established a list of users and

determine that the user is not the goal of data mining based on IP and Agent domain. From the test, we found that there are 28878 IPs and 1237 Agents and 32494 users, the average for each IP corresponds to 1.2 users. At the most one IP corresponds to 71 of the agent.

5. Session identification

If the time between two requests of the same user is more than 25 minutes and 30 seconds, a request is put as the beginning of another session, and the session information is recorded in the SESSION_MAIN in the session hits record keeping in the SESSION_LOG.

3.5 Preprocessing of the data warehouse

Having already discussed/introduced data loading, data clean-up and user identification, this section mainly discusses the transaction identification of algorithm to the path, followed by discussion of frequent access path followed by an analysis of the test data.

3.6 The maximum forward path algorithm

The maximum forward path for solving the user's browsing history is divided into sessions and, each session has its history. Here follows the process of requesting referrer domain data:

Click records into a user clicks on the event (Request, Referrer) format, then click on the event list (stored in the SESSION_LOG table).

Maximum forward path (Maximal of Forward Reference for MF) algorithm:

List MFSet = {MF1, MF2 ... MFn} saves a session prior to the results of the path, MFi (0 <i <n) for this session of maximum forward path.

The structure of the MF [C1, C2, C3, ... of Cn], where C1, C2, C3 ... of Cn for the user click events, is sorted by time.

The initial case MFSet is empty. to the click event in a user session, D1, D2, D3 ... Dn].

Clear the same click event (the referrer domain the request domain is the same).

For Di in D1, D2, D3, ... Dn]

Di of the above if MFSet

The [Di] joined MFSet

else
Di of the above, Ci (0 <i <n +1) in MFi, MFi for [C1, C2, C3, ... of Cn]
if i == n
The MFi = [C1, C2, C3 ... of Cn, Di]
else
[C1, C2, C3 ... of Ci, Di] joined MFSet

As mentioned in Chapter 2, the algorithm for calculating the maximum before the path contains the path to fill the whole process, simplifying the handling process.

3.7 Frequent Travel Path Found

In order tostore MF to the database, the following treatments are recommended:

An isolated MF sub-path, if the MF = [C1, C2, the C3 ... Cn] MF sub-path refers to all non-repetitive promoter sequence:

[C1], [C1, C2] ... [C1, C2, ... Cn], [C2] [C2, C3] ... [C2, the C3 ... Cn] ... [Cn-1], [Cn-1, Cn], [Cn].

Storage of this sub-path to the database, if there is the path recorded in the support column 1, if this record does not exist, insert the record, and initialize the support of 1.

This method can avoid the traditional practices that may result in frequent path loss for the requirements of the new processing session, and can easily add to the result of previous results . All records in the database are available through a simple database query results, after frequent path threshold.

3.8 Example Analysis

In order to more clearly illustrate our algorithm, let us consider a session record in the table below:

Table 2. User's session records

#	Request	Referrer
1	a.html	-
2	b.html	a.html
3	c.html	-
4	d.html	a.html
5	d.html	b.html
6	a.html	-
7	e.html	c.html
8	f.html	d.html
9	g.html	f.html
10	h.html	d.html

First we clear the same click event on the 1st and the 6th, , the record is clear on the 6th. The initialization list MFSet is empty.

For the 1st event (a.html, -), MFSet is above, so that the MF1 = [1], MFSet = {the MF1};

The events of the 2nd, the last element of the MF1 is above, so that the MF1 = [1,2];

The event on the 3rd MFSet is the above, so that the MF2 = [3] MFSet = {MF1, the MF2};

The events of the 4th, the elements of the MF1 are above, so that MF3 = {1,4}, MFSet = {MF1, the MF2, MF3};

The events of the 5th, the last element of the MF1 are above, so that the MF1 = [1,2,5];

The event on the 7th, the last element of the MF2 of the above, so that the MF2 = [3,7];

The events of the 8th, the last element of the MF1 are above, so that MF3 = [1,2,5,8];

The events of the 9th, the last element of the MF1 are above, so that MF3 = [1,2,5,8,9];

The events of the 10th, three elements of the MF1 are above, so that MF4 = [1,2,5,10];

This session user is set to the path MFSet = {the MF1, the MF2, MF3, MF4} = {[1,2,5], [3,7], [1,2,5,8,9], [1,2,5,10]}.

Save this MFSet set to the database: MF4, it is necessary to save all of its sub-MF sequence [1], [1,2], [1,2,5], [1,2,5,10], [2] [2,5], [2,5,10], [5], [5,10], [10], the relevant records in support plus 1, did not insert a record.

SUMMARY

As competition on the Internet in increasingly competitive, personalized service has become an important direction of development of Web applications, tapping the user's interest, to provide a reference for e-commerce market decisions. Therefore these urgent needs of web log mining have become a research hotspot.

This thesis first described the general practice of the Web log, log preprocessing, then the implementation of the preprocessing system for a web log. Two algorithms proposed in Chapter 3 is an interesting attemptfor model assumptions . However, a detailed theoretical derivation was not conducted because of the frequent path discovery algorithm complexity and massive system memory consumption when running a large number of IO operation. , In addition, this system cannot be automated to run the report analysis and it is also necessary to manually operate the database. For large-scale log analysis, there is need for further improvement.

REFERENCES

- M. Agosti, G.M. Di Nunzio and A. Niero“From Web Log Analysis to Web User Pro-filing” In DELOS Conference 2007. WorkingNotes. Pisa, Italy, 2007, pp 121–132. (Accessed 5 February 2012)

- B. Berendt, B. Mobasher, M. Nakagawa,M. Spiliopoulou “The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis”, WEBKDD 2002, LNAI 2703, pp 159-179, 2003. (Accessed 10 February 2012)

- C. W. Cleverdon “The Cranfield Tests on Index Languages Devices”. In Readings in Information Retrieval, Morgan Kaufmann Publisher, Inc., San Francisco, California, pp.47–60, 1997. (Accessed 15 February 2012)

- F.M. Facca, P.L. Lanzi “Mining interesting knowledge from Weblogs: a survey”, Data and Knowledge Engineering Vol. 53, No. 3,June 2005, pp 225-241. (Accessed 25 February 2012)

- P.M. Hallam-Baker, B. Behlendorf “Extended Log File Format, W3C Working Draft WD-logfile-960323” http://www.w3.org/TR/WD-logfile.html. (Accessed 3 March 2012)

- D. Nicholas, P. Huntington, A. Watkinson“Scholarly journal usage: the results of deeplog analysis”, (Accessed 19 March 2012)