The Research and Implementation of a Correlative Degree Mining Algorithm Based on IIS Logs

Lei-Yue YAO
Software Research Department, Jiang Xi Blue Sky University
leiyue_yao@163.com

Jian-Ying_Xiong
Computer and Information Department, Jiang Xi Blue Sky University
xjy9@21cn.com

Abstract

In order to find out the user patterns that hide in web logs, log mining technology is one of the best ways. Log mining is the usage of data mining in the field of web server' logs. Although there are a set of softwares which can be used to analysis web logs, the algorithm raised in this article pay special attention to discover the relationship among all the pages of the web site. In this algorithm, size-link radio and static inner-link degree was creative used. According to the result of experiment, this algorithm can exactly find out the correlative ones among massive pages.

Keywords: *Web usage mining, Web log analysis, IIS log analysis*

1. Introduction

With the rapid development of Internet, web-sites and web applications are getting more and more important in people's daily life. Web log mining is the technology to find out the hidden user patterns by analysis the information that include in web logs. So that capability, structure and individual functions can be optimized and provided by using this statistics.

There are three ways to record user visit information: web server, web proxy server and user client. In the web server logs, one web-site's visit information of multi users is recorded. In the web proxy server logs, multi web-sites' visit information of multi users is recorded. In the client logs, multi web-site's visit information of one user is recorded. During above three logs, the first and the second ones are recorded automatically, and the third one should be recorded by appropriative software. Generally speaking, web server log has the highest structured form.

In this paper, we pay special attention to the web server logs, through the research and analysis of web server logs, the pages which are outlying ostensibly, but related actually are discovered in order to optimized web-sites structure in a more reasonable way. And these mining statistics are finally used to establish fast and direct information channels.

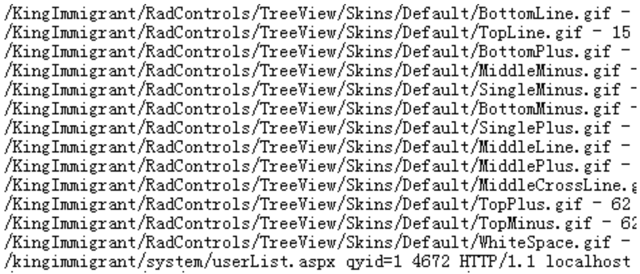
Log mining is a new field of computer information research. Log data preconditioning and mining algorithm actualizing are the only two common steps of log mining have reached the same idea. The algorithm mentioned in this paper also follows the common way.

2. IIS log preconditioning

Data preconditioning is distilling, disassembling, combining and transforming the row data into a structured form data, and then, insert these preconditioned data into database as the basal records which will be used to analyzed. Data preconditioning plays an important role in the data mining algorithms. Generally speaking, there are four steps, such as data cleaning, user detecting, colloquy detecting and path recruiting.

2.1. Data Cleaning

Although IIS log is highly structured and include sufficient information, there are also innumerable useless records exist in it. According to Fig. 1:



```
/KingImmigrant/RadControls/TreeView/Skins/Default/BottomLine.gif -
/KingImmigrant/RadControls/TreeView/Skins/Default/TopLine.gif - 15
/KingImmigrant/RadControls/TreeView/Skins/Default/BottomPlus.gif -
/KingImmigrant/RadControls/TreeView/Skins/Default/MiddleMinus.gif -
/KingImmigrant/RadControls/TreeView/Skins/Default/SingleMinus.gif -
/KingImmigrant/RadControls/TreeView/Skins/Default/BottomMinus.gif -
/KingImmigrant/RadControls/TreeView/Skins/Default/SinglePlus.gif -
/KingImmigrant/RadControls/TreeView/Skins/Default/MiddleLine.gif -
/KingImmigrant/RadControls/TreeView/Skins/Default/MiddlePlus.gif -
/KingImmigrant/RadControls/TreeView/Skins/Default/MiddleCrossLine.
/KingImmigrant/RadControls/TreeView/Skins/Default/TopPlus.gif - 62
/KingImmigrant/RadControls/TreeView/Skins/Default/TopMinus.gif - 6
/KingImmigrant/RadControls/TreeView/Skins/Default/WhiteSpace.gif -
/kingimmigrant/system/userList.aspx qyid=1 4672 HTTP/1.1 localhost
```

Fig. 1 a Segment of IIS

From above picture, we can easily point out that there is only one record is the visit information of web pages. And other records are useless. Therefore, the first thing should be tackled with is to filter the useless data from the log. Fig. 2 shows the flow path of the algorithm.

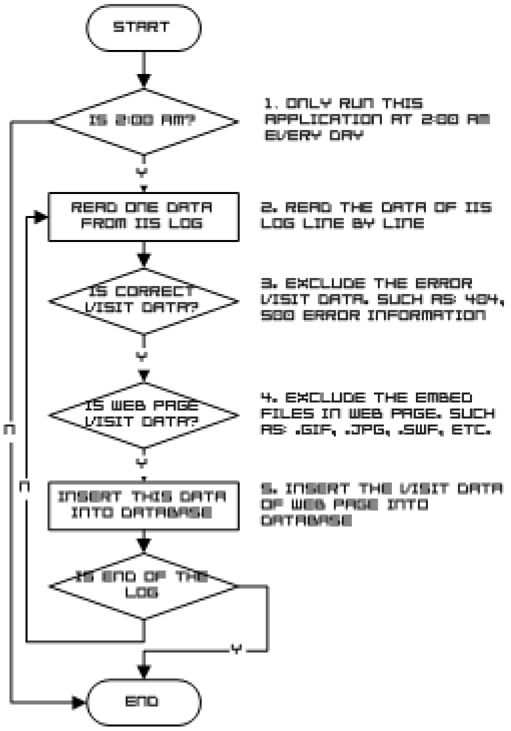

START
IS 2:00 AM?
1. ONLY RUN THIS APPLICATION AT 2:00 AM EVERY DAY
Y
READ ONE DATA FROM IIS LOG
2. READ THE DATA OF IIS LOG LINE BY LINE
IS CORRECT VISIT DATA?
3. EXCLUDE THE ERROR VISIT DATA. SUCH AS: 404, 500 ERROR INFORMATION
Y
IS WEB PAGE VISIT DATA?
4. EXCLUDE THE EMBED FILES IN WEB PAGE. SUCH AS: .GIF, .JPG, .SWF, ETC.
Y
N
N
INSERT THIS DATA INTO DATABASE
5. INSERT THE VISIT DATA OF WEB PAGE INTO DATABASE
IS END OF THE LOG
Y
END

Fig 2. The Flow Chat of Data Cleaning

2.2. User Detecting

One of the most important applications of correlative degree mining algorithm is finding out the favorite information of a certain user and displaying the data in the most easy-fetch way. So, finding out visit information of a certain user among the massive log data is a tough problem which should be tackled with in preparation. However, the variable Internet access patterns, such as local hash, proxy server and firewall, make great troubles in the field of user detect. Although Cookie and user register information are two efficient and effective ways to detect user identification, privacy and un-user-friendly are two fatal defects of the methods. In our algorithm, only can IP address, operation system, explore and the version of http protocol match, a certain user can be detected.

2.3. Colloquy Detecting

The purpose of colloquy detecting is to divide IIS log into several colloquy by a certain time slot. In other words, that is setting a time threshold to divide different colloquies. For example, consider Timeout as the time threshold, if record i and record j fit the following expression, they will be arranged into one colloquy. The default value of time threshold is 30 minutes.

$$Time_i \text{ - } Time_i <= T\ imeout$$

Time is still the only criterion to detect the valuable pages in the colloquy. As we all know, users always pay more time on their interested pages, and these pages are concerned to the algorithm. However, there is no "time cost" record in IIS logs. In order to solve this problem, "time cost" can be calculated as Definition 1.

Definition 1 Suppose there are two pages ($page_i$ and $page_{i+1}$) are continuously visited by a same user, T_i and T_{i+1} are the login time of $page_i$ and $page_{i+1}$, the time spent on $page_i$ is $T_{i+1} - T_i$.

If $T_{i+1} - T_i <= K$ (a threshold defined by programmer, its default value is 5 minutes), $Page_i$ is considered as the valuable page. Moreover, the last page in each colloquy is also considered as the valuable page.

3. Implementiation Of Correlative Mining Algorithm

Before introducing the algorithm, there are two important conceptions, "Weighted Size-Link Ratio" and "Static Inner-Link Degree", should be explicated.

3.1. Size-Link Radio

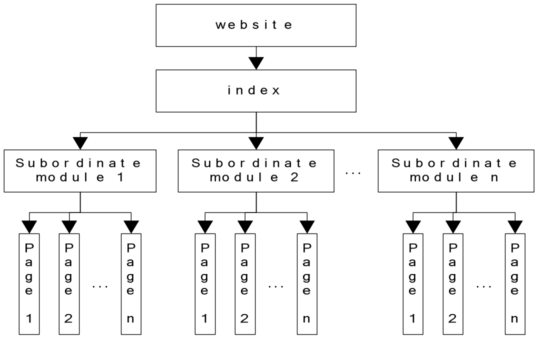


Fig 3. The Topology of a Common Web-site

According to Fig 3, it can be inferred that there are at least three kinds of pages in a web-site. Such as, index page, secondary page and content page. As we all know, once we enter a web-site, a long time is always spent on its index page to find out the information that we interested in. www.hao123.com is an example in certain. In this case, correlative degree must be not veracious enough if mining algorithm easily goes on the way that mentioned in section 2.3.

Different weight should be endowed with different kinds of pages. For example, content pages should have higher weight to index page. In order to quantize this concept, Weighted Size-Link Ratio is imported in our algorithm.

Definition 2 The ratio of page size and links exists in the page is called Size-Link Ratio (SLR).

Suppose "K" is the unit of page size, if a page's size is 3K which includes 20 links, the SLR of this page is 3/20. Theoretically, SLR is a number between 0 and infinitude. To fit the needs of our algorithm, SLR is mapped a number between 0 and 1. We call this mapped number as Weighted Size-Link Ratio.

Definition 3 Weighted Size-Link Ratio is a float between 0 and 1 which is mapped by SLR. The mapping formula is list as follow:

$$WSLR = (1 - e^{-SLR})$$

From above definitions, we can inferred that if a page's size is small and includes many links, WSLR of this page is small, and vice versa. Considering a especial situation that there is any link in a page, the above formula can be modified as follows:

$$\begin{cases} WSLR = (1 - e^{-SLR}) & \textit{link > 0} \\ WSLR = 1 & \textit{link = 0} \end{cases}$$

3.2. Static Inner-Link Degree

In order to build a user-friendly web-site, during the period of developing, web designer always put all the content pages which contain the same keywords into a certain group. The default relationship which is found by designers is considered as Static Inner-Link.

However, it is a more common phenomena that viewer always interested in several pages which have no Static Inner-Link at all. The relationship among this kind of pages is considered as Dynamic Inner-Link. A good mining algorithm is to find out all the Dynamic Inner-Link among web pages while make sure that the Static Inner-Link among the result pages is as low as possible. In order to quantize this concept, Dynamic Inner-Link Degree is imported in our algorithm.

Suppose the relationship among a set of pages in one colloquy is a directed graph, where pages are the nodes of the graph, and links among these pages are its border. If the collection of borders is null, means that there is no direct way to reach each other among these pages. In another extreme, there are enough direct ways to reach each other between every two pages. Static Inner-Link Degree is used to quantize the relationship of the pages.

Definition 4 The formula to calculate Static Inner-Link Degree(SILD) can be list as follows:

$$\begin{cases} SILD = Graph(G) / (G * (G - 1)) & \text{G>1} \\ SILD = 0 & \text{G=1} \end{cases}$$

Where Graph(G) stands for the borders of a direct graph; "G" stands for pages of this graph. If there is no links between any two pages, the value of SILD is 0, while in another extreme, SILD's value is 1. From the analysis of section 2.2 and 2.3, it's can be inferred that pages do not belong the same web site when SILD's value is 0. And page can be judged as index page when SILD's value is 1. Both 0 and 1 can be excluded since the results are not valuable pages.

3.3. The Implementation of Algorithm

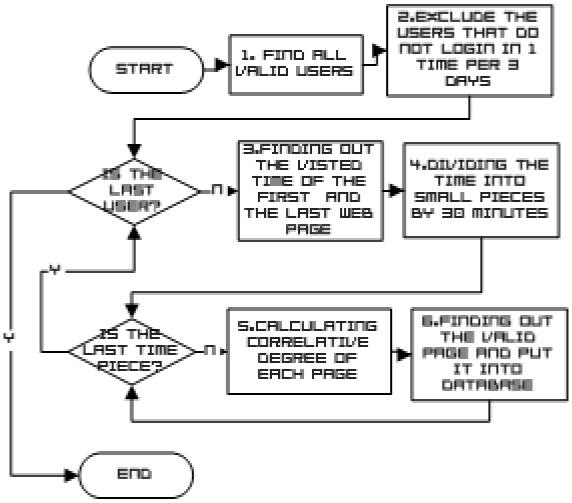

START
1. FIND ALL VALID USERS
2.EXCLUDE THE USERS THAT DO NOT LOGIN IN 1 TIME PER 3 DAYS
IS THE LAST USER?
n
3.FINDING OUT THE VISTED TIME OF THE FIRST AND THE LAST WEB PAGE
4.DIVIDING THE TIME INTO SMALL PIECES BY 30 MINUTES
y
y
IS THE LAST TIME PIECE?
n
5.CALCULATING CORRELATIVE DEGREE OF EACH PAGE
6.FINDING OUT THE VALID PAGE AND PUT IT INTO DATABASE
END

Fig 4. The Flow Chart of Algorithm

To avoid the high pressure of web server, this algorithm is only triggered in the middle night (about 2:00 am). According to Fig 4, there are six main steps in this algorithm. First, extract the true users from the massive log data. Since there are innumerable useless visit information is left by google and baidu. Therefore, the first step is to filter these visit information. Second, mine out the real viewer during these visitors. Third, find out the first and last visit time of a certain web site. Forth, divide the time period into small pieces (30 minutes per piece). Fifth, according to the formula: WSLR / SILD, calculate the correlative degree of each page in every time piece. Finally, compare the result to the threshold, then, insert the eligible page information into database.

4. Test result

This algorithm has already run on the web server (http://218.65.86.43). The environment and log information are list as follows:

Web application name: The reservoir information management
system of Jiang Xi Pro.
IIS log file range: 2009-5-15 ~ 2009-5-17
Records in these IIS log files: 173,567
Viewers recorded in log files: 77
Valuable pages in log files: 113
Time cost of the algorithm: 671,325 ms
Element of the server: CPU: Intel® Xeon® 1.6 8 cores; Memory: 8G, Operation System: Windows 2003.

Tack a certain user as the example whose IP is 61.178.172.82, operation system is Windows NT 5.0, explore is MSIE 6.0 and the version of http protocol is HTTP/1.1. The results of the algorithm are list in Table 1.

Table 1. Test Results

Page Address	P.S.
/kingimmigrant/population/famlyList.aspx	Population list
/kingimmigrant/project/ghList.aspx	Project list
/kingimmigrant/help/sp/8-ghsb.htm	Help file
/kingimmigrant/Login.aspx	Login page（useless page）

Form the statistics shows in Table 1. It’s easy to conclude that the algorithm designed in this paper is highly effective in finding out the hidden relationship among all pages that record in IIS log file. Although, “/kingimmigrant/Login.aspx” is a mis-judged page, it can be excluded by programmers. Therefore, the algorithm is feasible.

5. Conclusion

The algorithm elaborated in this article is aiming at finding out the hidden relationship among different pages. Compare to current log analysis applications, the results of this mining algorithm has greater advantages in structure optimizing and capability improving.

The further work in this field is to combine the algorithm and e-business together. So that favorite information can be pushed to the viewer by programs.

6.Reference

[1] S. Papadimitriou, A. Brockwell, and C. Faloutsos. Adaptive, unsupervised stream mining. VLDB Journal, 13(3):222–239, 2004.

[2] P. Smyth, D. Pregibon, and C. Faloutsos. Data-driven evolution of data mining algorithms. Communications of the ACM, 45(8):33–37, 2002.

[3] J. T. L. Wang, M. J. Zaki, H. Toivonen, and D. E. Shasha, editors. Data Mining in Bioinformatics. Springer, September 2004.

[4] M. J. Zaki and C.-T. Ho, editors. Large-Scale Parallel Data Mining. Springer, September 2002.

[5] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the Privacy Preserving Properties of Random Data Perturbation Techniques. In Proc. of the 2003 IEEE Intl. Conf. on Data Mining, pages 99–106, Melbourne, Florida, December 2003. IEEE Computer Society.

[6] C. Giannella, J. Han, J. Pei, X. Yan, and P. S. Yu. Mining Frequent Patterns in Data Streams at Multiple Time Granularities. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, editors, Next Generation Data Mining, pages 191–212. AAAI/MIT, 2003.