International Journal of Emerging Engineering Research and Technology
Volume 3, Issue 10, October 2015, PP 50-55
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Frequent Pattern Mining in Web Log Data using Apriori Algorithm

[1]S.VijayaKumar, [2]A.S.Kumaresan, [3]U.Jayalakshmi

[1]MCA Department, Priyadarshini Engineering College, Vaniyambadi, Tamilnadu, India
[2]Department of Computer Science, Priyadarshini Engineering College, Vaniyambadi, Tamilnadu, India
3 MCA Department, Priyadarshini Engineering College, Vaniyambadi, Tamilnadu, India

ABSTRACT

With the growing popularity of the World Wide Web (Web), large volumes of data are gathered automatically by Web servers and collected in access log files. Analysis of server access data can provide significant and useful information. In this paper, we address the problem of Web usage mining, i.e. mining user frequent patterns from one or more Web servers for finding relationships between data stored and pay particular attention to the interesting new patterns. We adapt a very efficient Apriori algorithm for matching interesting new patterns and applied support and confidence to calculate the measures of interesting patterns, to this particular context.

Keywords: Data Mining, Frequent Pattern, Web Mining, Apriori Algorithm, Support and Confidence

INTRODUCTION

The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified such that they better suit the demands of the Web. New approaches should be used which better fit the properties of Web data. Furthermore, not only data mining algorithms, but also artificial intelligence, information retrieval and natural language processing techniques can be used efficiently. Thus, Web mining has been developed into an autonomous research area.

The focus of this paper is to provide an overview how to use frequent pattern mining techniques for discovering different types of patterns in a Web log database. The three patterns to be searched are frequent itemsets, sequences and tree patterns. For each of the problem an algorithm was developed in order to discover the patterns efficiently. The frequent itemsets (frequent page sets) are discovered using the Itemset Code algorithm presented. The main advantage of the Itemset Code algorithm is that it discovers the small frequent itemsets in a very quick way, thus the task of discovering the longer ones is enhanced as well.

RELATED WORK

Association Rule used for Web Mining

In Web usage mining several data mining techniques can be used. Association rules are used in order to discover the pages which are visited together even if they are not directly connected, which can reveal associations between groups of users with specific interest. This information can be used for example for restructuring Web sites by adding links between those pages which are visited together. Association rules in Web logs are discovered. Sequence mining can be used for discover the Web pages which are accessed immediately after another. Using this knowledge the trends of the activity of the users can be determined and predictions to the next visited pages can be calculated.

Address for correspondence:

vijayviswak@gmail.com

Proposed Apriori Algorithm for Web Mining

The name of the Apriori algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset property which is that all nonempty subsets of a frequent itemset must also be frequent. The main idea is to find the frequent itemsets. The process of the algorithm is as follows.

Step1. Set the minimum support and confidence according to the user definition.

Step2. Construct the candidate 1-itemsets. And then generate the frequent 1-itemsets by pruning some candidate 1-itemsets if their support values are lower than the minimum support.

Step3. Join the frequent 1-itemsets with each other to construct the candidate 2-itemsets and prune some infrequent itemsets from the candidate 2-itemsets to create the frequent 2-itemsets.

Step4. Repeat the steps likewise step3 until no more candidate itemsets can be created.

Counting Occurrence Algorithm

Before support and confidence for each rules is determined, number of occurrence for each rules must be calculated. An algorithm for counting the number of occurrence is shown below:

```
count = 0
x=recordname
move record first
do while not eof()
if(currentrecord=x)then
count=count+1
end if
move next
end do
```

Algorithm for Support

Support measures how often the rules occur in database. To determine the support for each rules produced, several arguments have been identified in calculating the support such as Total Transaction in database and number of occurrences for each rules. The formula for support is shown below

Input :-

Total Transaction in DB

No. of occurrences each item {x, y}

$$\textit{Support} = \frac{\textit{Number of occurrences } \{x, y\}}{\textit{Total Transaction in DB}}$$

Algorithm for Confidence

Based on the examples, support measures how often the rules occur in database while confidence measures strength of the rules. Typically, large confidence values and a smaller support are used (Dunham, 2002). Formula for calculating the confidence value is shown below

Input :-

Total occurrence for item X

Total occurrence for item X and Y

$$\textit{Confidence} = \frac{\textit{Total occurrence for item X and Y}}{\textit{Total occurrence for item X}}$$

Problem

Let us consider the part of the access log file given in Table 2.1.. Accesses are stored for merely four visitors. Let us assume that the minimum support value is 50%, thus to be considered as frequent a sequence must be observed for at least two visitors. The only frequent sequences, embedded in the access log are the following:

Table2.1. *web log data*

IP Address	URL Accessed	Time
rres1,ne.wi.ac.uk	/api/java.io.Bufferedwriter,html /api/java.util.zip, CRc32.html /api/java.io.Bufferedwriter,html /java-tutorial/ui/animLoop.html /atm /logiciels,html\ /relnotes/deprecatedlist.html	01/Jan/1999 01/Jan/1999 02/Feb/1999 04/Feb/1999 18/Feb/1999 18/Feb/1999
Acasun.cekerd.edu	/perl/perlre.html /java.tutorial/animLoop.html /html4,0/struct/global,html /api/java.util.zip,CRC32,html /postgres/html-manual/query.html	11/Jan/1999 12/Jan/1999 29/Jan/1999 29/Jan/1999 29/Jan/1999
Acccs.francomedia.gc.ca	/java.tutorial/animLoop.html /apache/manual/misc/API.html /postgres/html-manual/query.html /perl/perlre.html /api/java.io.Bufferedwriter,html	05/Jan/1999 05/Jan/1999 05/Jan/1999 12/Feb/1999 12/Feb/1999
ach3.pharma.mcgill.ca	api/java.io.Bufferedwriter,html /java-tutorial/ui/animLoop.html /html4,0/struct/global,html /postgres/html-manual/query.html /relnotes/deprecatedlist.html	06/Feb/1999 06/Feb/1999 07/Feb/1999 07/Feb/1999 08/Feb/1999

The table 2.1 shows the web log data.

Table 2.2 shows the numbering of URL"s numbered for the convenience of further processing.

Table2.2. *Numbering URL's*

URL	NUMBER
/api/java.io.Bufferedwriter,html	1
/api/java.util.zip, CRc32.html	2
/java-tutorial/ui/animLoop.html	3
/atm /logiciels,html\	4
/relnotes/deprecatedlist.html	5
/perl/perlre.html	6
/html4,0/struct/global,html	7
/postgres/html-manual/query.html	8
/apache/manual/misc/API.html	9

Using the numbers assigned to URL"s table 2.3 shows the summary of web log data which is called as L1

Table2.3. *Summary of Web log Data L1*

Ip Address	URL Accessed
A	1,2,1,3,4,5
B	6,3,7,2,8
C	3,9,8,6,1
D	1,3,7,8,5

SOLUTION USING APRIORI ALGORITHM

Table 2.4 shows the calculation of L2 from L1 in which the not frequently accessed URL's are removed.

Table3.1. *Calculation of L2*

Scan database for count of each candidate Calculation of L2

URL	Support Count
1	4
2	2
3	4
4	1
5	2
6	2
7	2
8	3
9	1

→

URL	Support Count
1	4
2	2
3	4
5	2
6	2
7	2
8	3

Table 3.2 shows the frequently accessed URL's by Pair. The table shows only the support count >=2 called as L3

Table3.2. *Calculation of L3*

URL	Support Count
1,3	3
1,5	2
1,8	2
2,3	2
3,5	2
3,6	2
3.7	2
3,8	3
6,8	2
7,8	2

Table 3.3 shows the calculation of frequent three URL's. The table shows only the support count >=2 called as L4

Table3.3. *Calculation of L4*

URL	Support Count
1,3,5	2
1,3,8	2
3,6,8	2
3,2,8	2

Table 3.4 shows the calculation of support and confidence of L3 in which we come to the conclusion that our frequently accessed URL's are interesting patterns. If the confidence > support then the rules for X-> Y are interesting rules.

Table3.4. *Calculation of Support and Confidence for L3*

URL	Total Occurrences of X&Y	Total occurrences of X	Confidence	Support(Total transaction = 21)
1,3	3	4	0.75	0.14
1,5	2	4	0.5	0.10
1,8	2	4	0.5	0.10
2,3	2	2	1	0.10
3,5	2	4	0.5	0.10
3,6	2	4	0.5	0.10
3.7	2	4	0.5	0.10
3,8	3	4	0.75	0.14
6,8	2	2	1	0.10
7,8	2	2	1	0.10

Similarly Table 3.5 shows the calculation of support and confidence of L4 in which we come to the conclusion that our frequently accessed URL's are interesting patterns. If the confidence > support then the rules for X-> Y are interesting rules.

Table3.5. *Calculation of Support and Confidence for L4*

URL	Total Occurrences of X&Y	Total occurrences of X	Confidence	Support(Total transaction = 21)
<1,3>,5	2	3	0.67	0.10
<1,3>,8	2	3	0.67	0.10
<3,6>,8	2	2	1	0.10
<3,2>,8	2	2	1	0.10

The contribution of the paper is to introduce the process of web log mining, and to show how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's navigation behavior. The system administrator could make a decision from the result illustrated in order to improve or enhance the content, link, site navigation and facilities.

CONCLUSION

Web Usage Mining is an aspect of data mining that has received a lot of attention in recent year. Commercial companies as well as academic researchers have developed an extensive array of tools that perform several data mining algorithms on log files coming from web servers in order to identify user behavior on a particular web site. Performing this kind of investigation on the web site can provide information that can be used to better accommodate the user's needs.

REFERENCES

[1] Renáta Iváncsy, István Vajk, Frequent Pattern Mining in Web Log Data. In Acta Polytechnica Hungarica Vol. 3, No. 1, 2006

[2] Lee M L, Ling T W, Low W L. IntelliClean: a knowledge-based intelligent data cleaner [A]. In: Proceeding of the 6th ACM SIGKDD International Conference on Knowledge discovery and Data Mining[C]. Boston: ACM Press, 2000: 290 - 294

[3] Chen Wei, Ding Qiu-lin. Edit distance application in data cleaning and realization with Java[J].Computer and Information Technology, 2003,11(6):33 - 35

[4] Bunke H, Jiang X Y, Abegglen K,et al. On the weighted mean of a pair of strings [J]. Pattern Analysis & Applications, 2002,5(5): 23 - 30

[5] Chen Wei, Ding Qiu-lin, Xie Qiang.Interactive data migration system and its approximately-detecting efficiency optimization[J].Journal of South China University of Technology (Natural Science Edition), 2004, 22(2): 148 - 153

[6] Batista G E A P A, Monard M C. An analysis of four missing data treatment methods for supervised learning [J]. Applied Artificial Intelligence, 2003,17(5-6): 519 - 533

[7] Srikant R., Agrawal R., "Mining Generalized Association Rules", Proceedings of the st International Conference on Very Large Databases (VLDB'95), Zurich, Switzerland, September 1995, p. 407-419.

[8] Srikant R., Agrawal R., "Mining Sequential Patterns: Generalizations and Performance Improvements", Proceedings of the 5th International Conference on Extending Database Technology (EDBT'96), Avignon, France, September 1996, p. 3-17.

AUTHOR'S BIOGRAPHY

S.Vijaya Kumar had completed his MCA from University of Madras in the year 2001, and had completed his M.Phil in the year 2007. He had qualified for SET (State Eligibility Test for Lectureship) in 2012. He had authored 2 Research papers. He is currently working as Assistant Professor (Sr) in the Department of Computer Applications, Priyadashini Engineering College, Vaniyambadi, Vellore Dt, TN. His area of interest includes Parallel Processing, Embedded Systems, Artificial Intelligence and Data Mining.

A.S.Kumaresan had completed his BE from University of Madras in the year 2002, and had completed his ME in the year 2005. He had authored 1 Research paper. He is currently working as Associate Professor in the Department of Computer Science and Engineering, Priyadashini Engineering College, Vaniyambadi, Vellore Dt, TN. His area of interest includes Data Base Management System, Artificial Intelligence and Data Mining.

U.Jayalakshmi had completed her MCA from Anna University, Chennai in the year 2009. She is currently working as an Assistant Professor in the Department of Computer Applications, Priyadarshini Engineering College, Vaniyambadi, Vellore Dt, TN. Her area of interest is Data Structures, Design and Analysis of Algorithms and Data Mining.