TELKOMNIKA, Vol.16, No.5, October 2018, pp.2118~2126
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v16i5.7573

Website Content Analysis Using Clickstream Data and Apriori Algorithm

Supriyadi*[1], Yani Nurhadryani[2], Arif Imam Suroso[3]
[1]Informatics, STMIK Kharisma Karawang, Jl. Pangkal Perjuangan Km.1 Bypass Karawang 41316, Indonesia
[2]Department of Computer Science, Bogor Agricultural University, Kampus IPB Darmaga, Jl. Meranti, Wing 20 Level 5-6, Bogor 16680, Indonesia
[3]Department of Management, Bogor Agricultural University, Kampus IPB Darmaga, Jl. Meranti, Wing 20 Level 5-6, Bogor 16680, Indonesia
*Corresponding author, e-mail: fnfcreator@stmik-kharisma.ac.id[1], ynurhadryani@yahoo.com[2], imamsuroso@gmail.com[3]

Abstract

Clickstream analysis is the process of collecting, analyzing, and reporting data of visited pages by visitor at the time of mouse clicks. Clickstream data are generally stored on a web server in the access.log file including IP Address data, reference pages, and access time. This study aims to analyze clickstream data by converting into the form of a comma sparated value (csv) so that the string inside of it could be grouped and stored in a database. The important information in the database was processed and retrieved by using one of the techniques in web mining called apriori algorithm analysis. Apriori algorithm implementation was done at the time of reading the database and table query analysis on the software developed. Results of this study were the statistics describing the level of access to web pages that were very helpful for web developers to develop web sites.

Keywords*: clickstream data, website content, database analysis, apriori algorithm*

1. Introduction

Website visitor or users activities are varied and heterogeneous in terms of habits and access time. The user is the person trying to search something by typing, speaking or clicking into a web browser with a personal computer or mobile device. All activities are recorded by the Web Server and stored in the access log file. The file is recorded each time the user makes a change process click (clickstream) to link in a web page, which is generally called clickstream data. Clickstream data can be analyzed in a particular area such as a web page, client login, web server, router, or server proxy [1]. The key issue is that on the server side there is an aggregate picture of the usage of a service by all users, while on the client side there is complete picture of usage of all services by a particular client, with the proxy side being somewhere in the middle [2]. The data analysis techniques of clickstream can be done in several ways, such as by identifying unique users and transactions [3], modeling the behavior of the user in the form of a tree behavior of users [4,5] and reading the clickstream data using a computer programming language [6].

The process of analyzing the clickstream data is part of the Web Usage Mining (WUM) who performs a discovery data using a secondary data that is available on a web server, which includes data access logs, browser logs, user profiles, registration of data, user session, cookies, user queries and also mouse click data [7]. There are three important stages in data mining websites that need to be done [8,9], the first step is to clean up the data as an initial iteration and preparing to take the data patterns of usage by Web site users. Step two is to extract patterns of data usage that has been acquired, and step three is to build a predictive model based on the data that has been extracted earlier. The data cleaning stage is a stage that is the most in need of high resources because the amount of data that were cleared. The primary goal of a data cleaning effort is to eliminate data inconsistencies, invalid values, and other shortcomings in data integrity from the legacy databases.The main data cleaning processes are editing, validation and imputation. Fill in missing values, smooth noisy data,

Received October 14, 2017; Revised March 9, 2018; Accepted September 7, 2018

identify or remove outliers and noisy data, and resolve inconsistencies. Concrete data mining before the data are the following Web data filtering, Anti Internet spider, User identification, Session identification, and Path completion [10]. The number of data that is cleaned varied according to the needs of research and could reach 88.7% [11].

Websites that have solid data traffic will form a very large log file access. Due to the number of text data that is processed, it takes a lot of techniques to reduce data processing time of access log. The technique is carried out in the form of algorithms or can be modified by using parallel computing techniques [12]. Various computer applications were developed to read the log access file made to be easy to read as the Apache log viewer that is owned by Apache web server, Webalizer and AWStats. The data presented in the application in a general new form of grouping data based on elements such as the log is based on the IP address, time of access or the most accessed pages. While web developers require additional information such as information on where the web page is accessed or information value of connections between web pages as a reference in the maintenance and development of the website. A glimpse of these problems can be overcome by performing data clustering techniques to the data log access referrers as was done in the techniques of clustering the search engine automatically [13] but the data referrers come from different sources, namely from search engines and another web address.

The study tries to provide an alternative solution for managing clickstream data with a database management approach. The approach is done by algorithms integration of priori and structured query language (SQL) in a web-based computer application. These applications are designed to be able to perform pre-processing process that includes a cleaning process of clickstream data and analyzing the relationship between web pages using a common analysis of the a priori algorithm that analyzes a shopping cart that is used to generate association rules [14]. Similar analyzes were performed by Latheefa [15] in processing the clickstream data, but the applications developed by its emphasis on connections between web pages are accessed by folder. While this research's connections on the analysis web page is generated based on the files in the folder.

2. Research Method

The study had been conducted by taking secondary data from the Website of Indonesian Ministry of Agriculture or the Ministry of Agriculture of the Republic of Indonesia (MOA) by using time interval of log server for two months, i.e. November 2012 to December 2012. The selection of the data was only as samples to be analyzed for the development of software that could process access log data for any time period. Generally, this research was done by the following the three main stages as shown in Figure 1 [16].

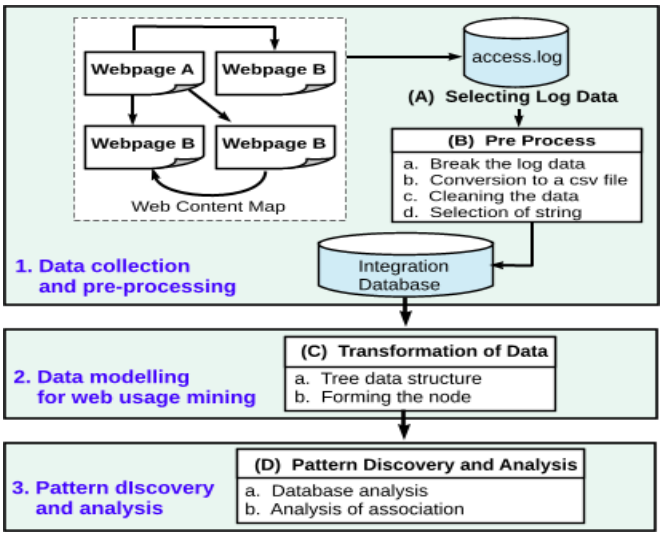

Figure 1. Three main stages of research

2.1. Selecting Log Data

String log follows the format of the log on the Apache web server using the following explanation. The explanation of string log web server apache format as shown in Table 1.

LogFormat: "%h %l %u %t \"%r\" %>s %b"

Table 1. The Explanation of String Log Web Server Apache Format

No	String	Description	Remarks
1	%h	IP Address *client* or remote *host* requesting the service to a host server	host
2	%l	Process of log user or client identification	log identification
3	%u	User id of visitors based on HTTP authentication	User id
4	%t	Time when request is received by the server	time
5	\"%r\"	Inline requests written by client	Request line
6	%>s	Status code given by *server* to *client*	status
7	%b	Data size or document given to client	byte

2.2. Pre Process

2.2.1. Break the Log Data

Data of access log is a text file with a very large size, especially if the website analyzed has sufficiently high number of transactions. The number of text data that causes the time to open the data is very long. Moreover, some text editors can not open access.log file. Technique employed here is a way to open a text file first overall with a special editor, then for some specific lines, the file is cut and moved to the new file to be saved.

2.2.2. Conversion to a Comma Separated Value (CSV) file

In order to process the data more flexible, it is first converted into a form of csv because this format can be converted to other forms, such as SQL format or spreadsheet.

2.2.3. Cleaning the data

Data to be cleared is the log data that has been broken and has been in the form of csv. The characteristic from csv file is only one string delimiter (sparator) and a cover string in every field. The separated data by sparator will be an array.

2.2.4. Selection of String

This study focused on the process of tracking the frequency of visits to any existing Web pages on a web page. From the log string format above; however, not all the block strings are taken, only a few are used in accordance with the needs of this research. Those which are used as an ingredient are:

LogFormat: “%h %t \ "%r\" %>s “

The retrival of four-group string above was based on the purposes of data to be processed, namely:

a. %h is a group of string describing the host that accesses the web server. The identity noted is address host or or IP Address. This data is very useful to know who accesses web pages.
b. %t is the time series performed by each host in the session log. These data are very useful to form one series of graph per each log for each host.
c. %r is a set of strings that contains the data transfer method (POST/GET) and request of web page by the user, these data will be used as the basic material to be used as a node in the graph series.
d. %s is the status generated by the protocol of HTTP (Hyper Text Transfer Protocol) regarding the success or failure of the communication process between the service requester (client) and web service providers.

2.2.5. Database Integration

After all the data is prepared in csv format, it is necessary to synchronize with the database that will be designed so that the log data can be integrated or imported into the

database. In general, the content of the tables in the database includes three groups:
a. The raw data or preliminary data
b. The data that has been cleaned
c. The results of data processing using apriori algorithm.

2.3. Transformation of Data

2.3.1. Tree Data Structure

Columns of data as a starting material of node formation are column line request. From the data request can be taken some further information that can be used as a node in the tree data structure. For the formation of the tree data structure, it is needed to limit data to be processed so that not all data is included as follows:
a. Limitation of time series in this case the time is limited to the range by one day, assuming that when the day changes, the route will be renewed again.
b. The request data used are successfully processed by the web server with the status code of 200 (the process of communication with the server succeeds).
c. Data sent by GET method are excluded because this study did not discuss until all elements of the request (query) but only at the level of access to the page; therefore, GET and POST methods are considered equal.

If it is decribed in the form of structure of tree data, the content of Table 2 will be the tree structure as follows:

Table 2. The Example of file set accessed by host

host (%h)	File directory
157.55.33.40	/news/detailarsip_2.php
157.56.229.23	/news/detailevent.php
173.44.37.226	/wap/index.php
173.44.37.226	/wap/bpsdm/spp-kupang/index.php
66.249.77.63	/keuangan-perlengkapan/galeri.php
66.249.77.63	/pusdatin/statistik/metodologi/ tar.pdf

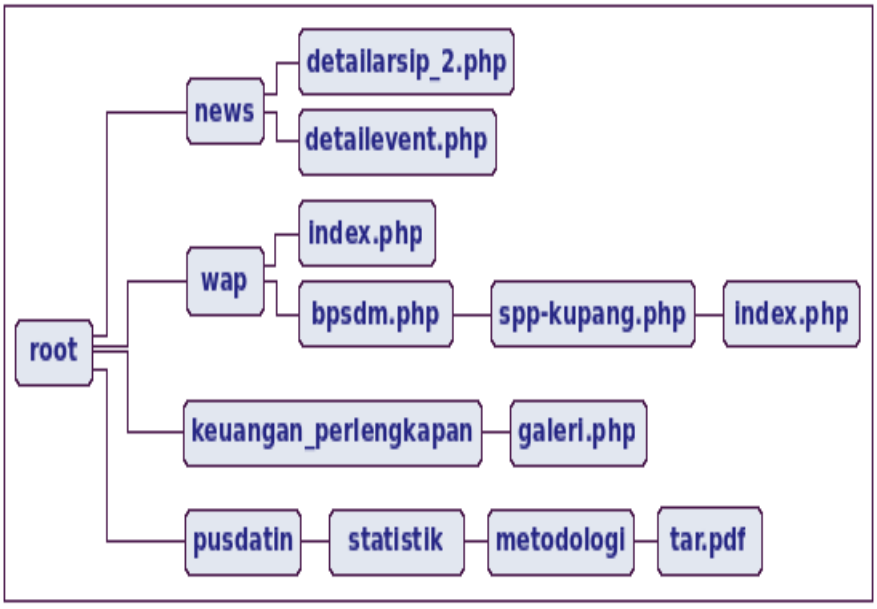

Figure 2. Web directory representation to the shape of tree structure

2.4. Forming the Node

To make the data process easy, it is needed to make unique index for every string of file directory series which is then called node. The example of file directory structure coding as

node as shown in Table 3. The example of accessdistribution frequency of web page as shown in Table 4.

Table 3. The Example of File Directory Structure Coding as Node

Coding	File Directory
A	/wap/bpsdm/spp-kupang/index.php
B	/news/detailevent.php
C	/news/detailarsip_2.php
D	/pusdatin/statistik/metodologi/ tar.pdf
E	/wap/index.php
F	/keuangan-perlengkapan/galeri.php

Table 4. The Example of Accessdistribution Frequency of Web Page

No	host (%h)	Node	A	B	C	D	E	F
1	157.55.33.40	C	0	0	1	0	0	0
2	157.56.229.23	B	0	1	0	0	0	0
3	173.44.37.226	E, A	1	0	0	0	1	0
4	66.249.77.63	F, D, B	0	1	0	1	0	1
	Total		1	2	1	1	1	0

2.5. Pattern Discovery and Analysis

2.5.1. Database Analysis

If string log data are converted into the form of csv and separation takes place by dividing empty space, it will look like the Table 5. It can be seen clearly that the group of string is accomodated in one coloum, unless the group of string %t and \"%r\" is accomodated in more than one coloum. This will affect the fields design made in log table. At first, the table has not adhered to the database regulation like primary key or index. This is done in order to make all data recorded first in the form of table to simplify the process of query.

Table 5. The Separation of String with Whitespace

1	2	3	4	5	6	7	8	9	10	11
66.249.73.7	-	-	[04/Nov/2012:04:09:51	+0700]	GET	/wap/index.php	HTTP/1.1	200	413	-
180.76.5.136	-	-	[04/Nov/2012:04:09:59	+0700]	GET	/news/index.php	HTTP/1.1	200	181	-

2.6. Analysis of Association

Association analysis is an analysis of the connection of web pages visited by the user. The technique used is to use the shopping cart analysis by using apriori algorithm. Through data analysis can enable businesses to grasp real-time market dynamics, optimize the O2O business platform, improve customer satisfaction, so that the business according to customer needs to develop a personalized, economic services, stable customer relationship. Apriori is one of the popular methods for discovering of knowledge discovery about finding the relationships among the items [17]. Aim of traditional association rule mining (Apriori) is to discover the frequent itemsets, which defines the itemsets of each transaction in the transactional database [18]. The object used as itemset is directory in the website or hereafter known as nodes. The directory consists of sub domains and folders in which in this study, it is considered as the same address.

The important information in the database was processed and retrieved by using apriori algorithm analysis. Apriori algorithm implementation was done at the time of reading the database and table query analysis on the software developed. Itemset is a set of web pages that are recorded on a data log and symbolized by $I = \{I_1, I_2, I_3, ... I_n\}$. While the transaction is a set of n transaction N symbolized by T. According to the association rules $X \rightarrow Y$ is a chance of a particular item appear together where X and Y are itemsets. To determine the value of the support is done by calculating the ratio of the number of total transactions itemset with the following formula:

$$Supp(x) = \frac{T(x)}{\sum_{j=1}^{n}(Tj)}; \quad Tx = transaction\ of\ x;\ x \in I;\ j = 1,2,3....,n;$$

The rules of association calculated that the probability of their confidence value itemset X and Y are itemsets in a transaction primarily to the following formula:

$$Conf(X \rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)}$$

3. Results and Discussion

3.1. Selecting data log

Log data recorded by the Apache web server are access.log, error.log and other_vhost_access.log. The selected log data in this study were data on the Website access.log Indonesian Ministry of Agriculture in the period of 2 months with a size of 632146266 bytes (2323844 lines of log).

The splitting of log data and the conversion into a csv file the log data that have been elected and then broken down to make it easier for the reading by program. The splitting of log data were not like dividing the data equally by using mathematical formulas, but this was taken by a certain number of lines. From the results of splitting, it was obtained the Table 6.

Table 6. Data of splitting results of access.log file

Number	Name of log file	Size (byte)	Number of rows
1	access_log_a.csv	110542665	480000
2	access_log_b.csv	150217973	522838
3	access_log_c.csv	147894289	475826
4	access_log_d.csv	93766392	311630
5	access_log_e.csv	62261364	253348
6	access_log_f.csv	67463583	280202
	Total	**632146266**	**2323844**

3.2. Cleaning the Data and String Selection

After the log data splitted into 6 groups, the 6 groups of data were imported into the database one by one like the following stages by using application developed based on web programming. There are 4 stages of cleaning as shown in Table 7 done as follows:

1. The first stage of cleaning is the process of text data entry of csv to the *data _log_1* table. In this stage, there is only import process. The cleaning is only replacing the quotation mark with blank space.
2. The second stage is the process of separation string request on the *data_log_1* by using script PHP.
   For example, the string request :
   /wap/index.php?option=component&id=&gbfrom=16258.
   On the data of request mentioned, all the strings are behind the question mark (?) which will be rased using the following command:
   $str_req = /wap/index.php?option=component&id=3&gbfrom=16258
   $split_req = explode('?',$str_req);
   After the command, there will be split array _request as the following:
   $split_req[0] = /wap/index.php
   $split_req[1] = option=component&id=3&gbfrom=16258
   After the data of *request* were divided into two, what is needed next is only the selection of string, namely *$split_req[0],* while *$split_req[1]* erased. After that the three letters on the right from *$split_req[0]* are kept on the fileds type_req and they will be the reference of query. The cleaning of data rows containing file of picture, audio, video, layout web and string query will be done by using the SQL command as follows:
   $pj_string = strlen($split_request[0]);

$type_req = substr($split_request[0],$pj_string-3,3);
$bersihkan = mysql_query("delete from data_log_1 where type_req ='css' or type_req ='js' or type_req ='.js' or type_req ='.db' or type_req ='xml' or type_req ='bmp' or type_req ='gif' or type_req ='jpg' or type_req ='jpeg' or type_req ='png' or type_req ='rc=' or type_req ='MYI' ortype_req like '%/' ");

3. The third stage of cleaning is the same process of request removal, and it is done at the same time in order to avoid data duplication.
4. The fourth stage of cleaning is a process of removal the transaction data done by the host (IP Address) at the same day with the same node access.

Based on data cleansing phase of table log access' number of transactions that are not required can be reduced up to 76 %, it is not much different from the results of research conducted by Latheefa [13], which managed to reduce the file size by 84% and while Kharwar [9] managed to reduce the file by 88.7%.

Table 7. The Stage of Access.Log Data Cleaning

No	Type of Transaction	Initial file	Result file	Number of transaction	Number of IP Host
1	First stage of cleaning	access_log.csv	data_log_1	2 323 844	34 060
2	Second stage of cleaning	data_log_1	data_log_2	1 641 919	34 060
3	Third stage of cleaning	data_log_2	data_log_1itemset	360 926	29 452
4	Fourth stage of cleaning	data_log_1itemset	data_log_1itemset	115 569	29 452

3.3. Tree Data Structure and Forming Node

Tree data structure that is formed is not represented in the form of images, but in the form of the directory tree. from the search results of MOA web directory, it can be gained that there are 20924 nodes or directories that are differenct each other by data sorting alphabetically *contain_node* in order to simplify the search.

3.4. Analysis of Database

The table of *data_log_1* is used to hold the log data when first read by the system, while *data_log_2* is a table to accommodate the primary log data for the data mining process. *data_log_2* table is used to determine the sequence of time within the limits of its smallest, the second, while *data_log_1* of itemset are used to form *ip_hos*t as the identification of per log session transactions with the most limited time, day or date. The following is the schema of table formation of *data_log_1*, *data_log_2* and *data_log_1itemset* as shown in Figure 3.

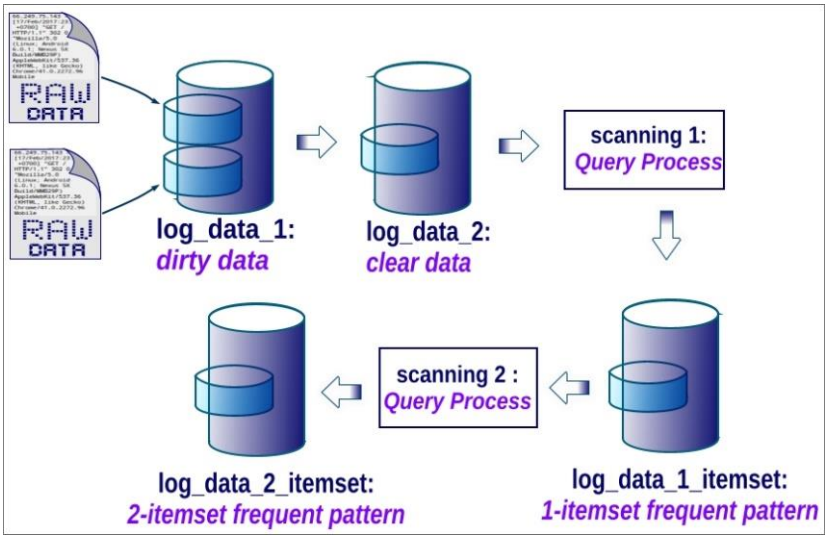

Figure 3. The scheme of clean log table formation

The following table data structures are used:

1. log_data_1={url_root, host, log_id, user_id, time, gmt, method, request, type_req, http_stat, status, byte, referer, User_Agent, UA_os, UA_ver, UA_lang, UA_1, UA_2, UA_3 UA_4, UA_5, UA_6, UA_7, UA_8, UA_9, UA_10 UA_11, UA_12, UA_13, UA_14, UA_15, UA_16}
2. log_data_2={url_root, ip_host, time, activity_node, type_req, description}
3. log_data_1itemset={url_root, ip_host, date, activity_node}
4. log_data_2itemset={url_root, ip_host, date, activity_node}

3.5. Analysis of Association

The first process of analysis conducted is by determining the candidate of 1-itemset for 8 highest transaction of node transaction on *data_log_1itemset* table with value of minimum support (MS) of 1% and the value of the Minimum Confidence (MC) of 0.2% can be seen in Table 8. After the first scanning is performed, it can be obtained the overall number of transactions of access node, 115569 transactions. It can be seen in support value that */index1.php* value is 8.8% meaning that 8.8% of all transactions contain 8.8% */index1.php* node, and so on for the other node data, the support value can be obtained with the same calculation. Determining candidate 2-itemsets can be done by searching the whole combination of access nodes contained in the 1-itemsets scan results, as shown in Table 8.

Table 8. The Scan Result of First Candidate 1-itemset

No	Node Code	Node Contents	*Node Count*	*Support (%)*	Fulfill the value of *MS*
1	8348	/index1.php	10162	8.8	Yes
2	8039	/event.php	8841	7.7	Yes
3	19262	/respon.php	8128	7.0	Yes
4	12223	/news/cover_es.htm	7097	6.1	Yes
5	20890	/wap/index.php	2808	2.4	Yes
6	19803	/tampil.php	2505	2.2	Yes
7	7530	/dir-alamatskpd/tampil.php	1961	1.7	Yes
8	10759	/news/detail.php	1496	1.3	Yes

Based on Table 8, the calculation process of confidence is then performed from association rules that qualify Minimum Support (MS) 1.0% and the Minimum Confidence (MC) 0.2% as follows. Support and confidence calculation results on associated 2-itemsets as shown in Figure 4.

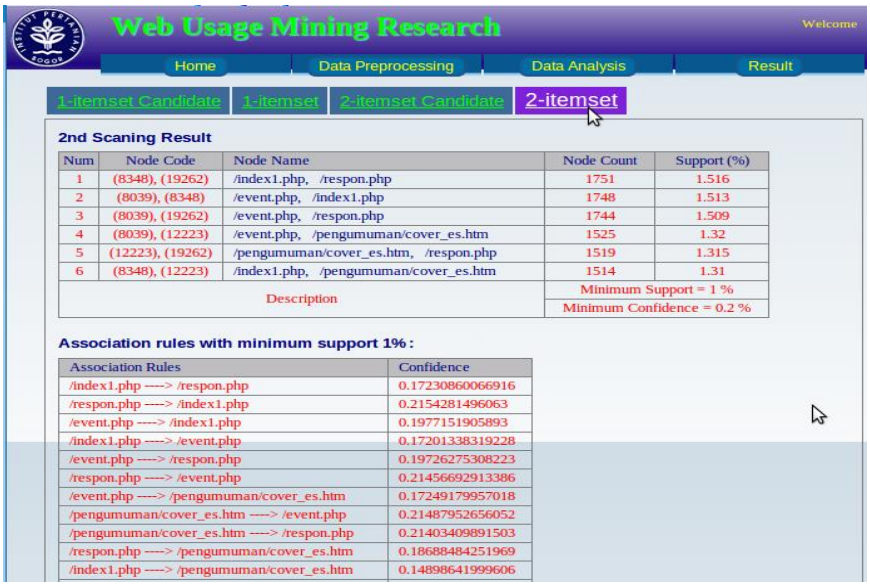

Figure 4. Support and confidence calculation results on associated 2-itemsets

4. Conclusion

This study had succeeded to make the process of log data preprocessing by using the web-based software developed and stored in a MySQL DBMS. By using the shopping cart analysis and minimum support and confidence which was set at 1% and 0.2%, the data obtained are as follows:

a. The most frequently accessed node or web page is /index1.php (table 8, which is the main page of the Ministry of Agriculture Web site. This shows that in tracing every sub domain or web page that exists on the MOA web, generally it must first pass the main page. Although */index1.php* is the most frequently accessed page, it does not reveal that the page is the most interesting page because */index1.php* is its default home page.
b. In the process of the 2nd scan, it can be obtained appropriate seven rules of associations. To develop the content related to the links, link suggestion can be put in the pages that meet the rule to a page that has a low hits.
c. If it is viewed from the average small value of support and confidence and the highest value of its support of around 8%, it can be said that the Ministry of Agriculture web site does not have a page that stands out the most accessible, meaning that traffic access to each page is relatively equal.

The technique of data readout in this study emphasis more on the process of database query, even though it seems slow but very effective to save all log data for each group of its string. Another alternative for reading the log data is by using the parser technique; however, it needs adjustment in its algorithm

References

[1] Moe WW, Fader PS. Capturing Evolving Visit Behavior in Clickstream Data. *Journal of Interactive Marketing*. 2004.
[2] Srivastava J, Desikan P, Kumar V. Web Mining - Accomplishments and Future Directions. Computer Science Department, University of Minnesota, Minneapolis. 2002. http://www.ieee.org.ar/downloads/Srivastava-tut-paper.pdf
[3] Cooley R, Mobasher B, Srivastava J. Data Preparation for Mining World Wide Web Browsing Patterns. *Journal of Knowledge and Information System*. 1999.
[4] Gunduz S, Ozsu MT. A Web Page Prediction Model Based on clickstream Tree Representation of User Behavior. *ACM*. 2003.
[5] Montgomery AL. Using Clickstream Data to Predict WWW Usage. Carnegie Mellon University. School of Industrial Administration. Pittsburgh, PA 15213-3890. 1999.
[6] Dinucă. An Application for clickstream analysis. *International Journal of Computers and Communications*. 2012; 6(1).
[7] Abdurrahman, Trilaksono BR, Mandala R. *Pemodelan Web Usage Mining untuk mengelola e-commerce*. Prosiding Konferensi Nasional Teknologi Informasi & Komunikasi untuk Indonesia. Institut Teknologi Bandung. 2006.
[8] Srivastava, Cooley R, Deshpande M, Tan PN. Web usage mining: Discovery and application of usage patterns from web data. ACM SIGKDD Explorations. 2000.
[9] Srivastava, Desikan P, Kumar V. Web Mining - Concepts, Applications and Research Directions. Department of Computer Science University of Minnesota, Minneapolis, MN 55455. 2005.
[10] Chongwen W, Scholten D. O2O E-Commerce Data Mining in Big Data Era. *TELKOMNIKA Telecommunication Computing Electronics and Control.* 2016; 14(2A).
[11] Kharwar AR, Naik CA and Desai NK. A Complete Pre Processing Method for Web Usage Mining. *International Journal of Emerging Technology and Advanced Engineering*. 2013; 3(10).
[12] Tong W, Pi-lian HE. Web Log Mining by an Improved AprioriAll Algorithm. *PWASET*. 2005; 4.
[13] Martiana E, Rosyid N, Agusetia U. Mesin Pencari Dokumen Dengan Pengklasteran Otomatis. *TELKOMNIKA: Telecommunication Computing Electronics and Control.* 2010; 8(1).
[14] Goswami, Anshu C, Raghuvanshi. An algorithm for frequent pattern mining based on apriori. *IJCSE*. 2010; 2(4): 942-947.
[15] Latheefa V, Rohini V. Web Mining Patterns Discovery and Analysis Using Custom-Built Apriori Algorithm. *International Journal of Engineering Inventions*. 2013; 2(5): 16-21.
[16] Mobasher, Bamshad. Data mining for web personalization. The adaptive web, Springer Berlin Heidelberg. 2007.
[17] Chongwen W, Scholten D. 2O2O E-Commerce Data Mining in Big Data Era. *TELKOMNIKA: Telecommunication Computing Electronics and Control.* 2016; 14(2A): 396-402.
[18] Prasad KR. Optimized High-Utility Itemsets Mining for Effective Association Mining Paper. *International Journal of Electrical and Computer Engineering (IJECE).* 2017; 7(5).