I.J. Information Technology and Computer Science, **2018, 12, 21-26**
Published Online December 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2018.12.03

ICT Training Recommendation using Web Usage Mining

Susi Maulidiah
Department of Computer Science, Bogor Agricultural University, Bogor, 16680, Indonesia
E-mail: smaulidiah@yahoo.com

Imas S. Sitanggang and Heru Sukoco
Department of Computer Science, Bogor Agricultural University, Bogor, 16680, Indonesia
E-mail: imas.sitanggang@ipb.ac.id, hsrkom@ipb.ac.id

Received: 01 March 2018; Accepted: 15 September 2018; Published: 08 December 2018

***Abstract*—**The sustainability of a course and training institute depends on the availability of students. There are many ways to promote the courses and training programs including promoting it through the institution's website. The visitor behavior of a website have hidden information that can be found using web usage mining approach. This study aims to discover the hidden information from the visitor patterns of course website. The data used are web access log data of August 2016. Web usage mining process was done using the Co-Occurence Map Sequential Pattern Mining using Bitmap Representation (CM-SPAM) algorithm which is available in the SPMF tool. Based on sequential pattern mining on the access log data, this study recommends improvements regarding the website structure and information that should be displayed on certain web pages. This study also found that the visitors of course website interested in three page types: one day seminar, tutorial and the training program.

***Index Terms*—**CM-SPAM, recommendation, sequential pattern mining, web usage mining.

I. Introduction

Non-formal education units in Indonesia consist of courses and training institutions or *Lembaga Kursus dan Pelatihan* (LKP), study groups, community learning center, majelis taklim and other similar educational institutions. LKP is an educational institution that provides certain training which is required by individuals such as students to help their education process, graduate students to help them in looking for a job, or companies to improve their employees skills in order to increase the companies or institutions's performance.

As the use of internet access increasing, learning process no longer can only be done by face-to-face or offline mode. Numerous tutorials either in writing or e-book or through video can be accessed easily by individuals or institutions who want to increase their skills. This is one of the causes that decrease the number of learners in LKP, especially which is engaged in Information and Communication Technology (ICT). On the other hand, number of learners is one of the main conditions for the sustainability of LKP.

At present, there are 5902 LKPs in the field of ICT in Indonesia. LKP XYZ is one of them which was established in 2002. In addition to provide ICT training, LKP XYZ collaborates with International certification institutions such as Microsoft and Cisco. It also becomes a place of competency test of ICT Competency Certification Institution. Nowadays, the information of training, programs, courses, workshops organized by LKP XYZ is provided through various media such as flyers, brochures, banners, social media and LKP XYZ's website. The traffic statistics from the website show that there are 195 unique visitors and 9716 hits a day in 2016.

Data mining is very useful in education. The most popular techniques for data mining are clustering, classification, sequential pattern, prediction, and association rule analysis [1]. Educational data mining has used to make an assessment of the role of student gender on successive rates of educational completion in Australia [2]. A related study also has been done by applying the Growing Hierarchical Individual Growing Map algorithm to analyze a citizen portal [3].

Association rule mining was also applied in determining book recommendation based on book associations [4,5]. A book recommendation system for digital library based on the user's profiles was developed using association rule mining [6]. The sequential pattern mining using the AprioriAll algorithm was also applied to generate sequence patterns on the library transaction dataset in Bogor Agricultural University, Indonesia [7]. The results are frequent sequential patterns containing book sequences borrowed by students generated at minimum supports of 0.3, 0.2, 0.15 and 0.1 [7]. A data mining tool was developed using association rule mining and collaborative filtering that can be used by instructors to get recommendations in e-learning courses improvement [8]. The three classification algorithms namely Naïve Bayes, Bagging, and C4.5 were applied on the dataset of non-active students in Indonesia Open University (IOU) for the period of 2004 to 2012. The

results show that the Bagging method provides a higher accuracy than Naïve Bayes and C4.5 [9]. The book recommender system was proposed that mines frequently hidden and useful patterns from the book library records [10]. The system provides recommendations based on the pattern generated using associated rule mining technique [10]. The association rule mining using the ECLAT algorithm was conducted to extract frequent association terms on the final project text document of Bachelor Program on Computer Science, Bogor Agricultural University, Indonesia. The association patterns were generated at the minimum support of 0.1, 0.15, 0.2, 0.25, 0.3, and 0.35. The text documents are then grouped using the k-Means clustering algorithm with number of cluster (k) of 10 [11].

In addition, web usage mining was implemented to analyze self-help services for elderly people in Taiwan [12]. Sequential pattern mining can be applying in web usage mining using the CM-SPAM and CM-SPADE algorithms. CM-SPAM has better performance compared with CM-SPADE algorithm for web click stream data [13]. A study on weblog files from the Social Explorer online-GIS website uses sequential pattern mining with the CM-SPADE algorithm to analyze the frequent sequences generated during the online-GIS website visits [14].

To further investigate the website of LKP XYZ, the site visitor activities can be found using web mining approach. The objective of this study is to conduct web usage mining on visitor access logs of the LKP XYZ webpage using sequential pattern mining with the CM-SPAM algorithm to get website visit patterns. The results will be used in formulating the recommendation for LKP XYZ manager and the web developer.

II. Methodology

A. Data

Website visitor information is recorded in access log data. The access log dataset is a text file in NCSA combined log format. This format has the following fields: host, rfc931, username, date:time, request, statuscode, byte, referrer, user_agent, cookies. This study uses access log data from the LKP XYZ's website. The HTTP request consists of 3 parts: the requested resource (URL), the HTTP method and HTTP protocol. Information of which web page visited can be found in the URL part. A visitor is defined as one IP that visits the LKP XYZ's website in a day. One session was calculated as one visit or access to one web page, so that one visit can consist of several sessions. The session duration is counted with minimum 2 second/page [15]. If the users' access has the session less than 2 second/page then this access will be removed.

B. Research steps

This study was conducted in two phases, namely data preparation and pattern recognition which is illustrated in Fig. 1. In the data preparation phase, web traffic is analyzed to obtain the data, then data acquisition and data preprocessing of the access log were performed to generate the web visitor log dataset. In the pattern recognition phase, sequential pattern mining was implemented to find patterns from the dataset. The patterns will be further analyzed to formulate the recommendation for LKP XYZ.

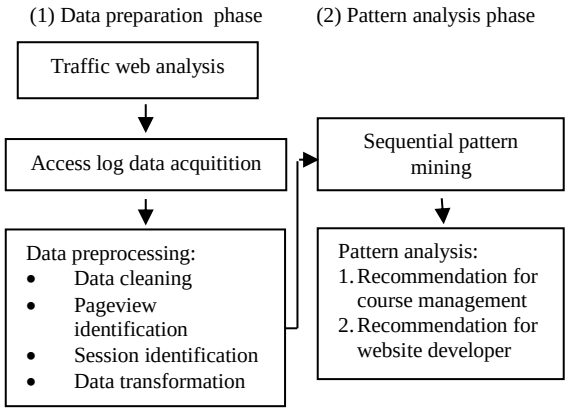

Fig.1. Research steps.

III. Data Preparation Phase

Web access log data need to be preprocessed and analyzed in order to meet the data format for further analysis. In this phase there are several steps to be conducted including web traffic analysis, access log data acquisition, and data preprocessing.

A. Web Traffic Analysis and Access Log Data Acquisition

Analysis of web traffic was performed using the access log file in August 2016. The data in August 2016 is the largest file among other log files meaning that in this month the LKP XYZ's website was frequently accessed by visitors. In addition, August is a holiday period for students in elementary school, junior and senior high school. In this period, graduate students find a job and add ICT skills to become more valuable candidate in job markets. The data were adjusted based on the format for further analysis in web mining. The access log data acquisition was done using Java Programming.

B. Data Preprocessing

Data preprocessing steps were performed in the access log file. The steps include:

- *Data Cleaning*

The data with the "user_agent" field which contain "bot" and "spider" values were deleted. Web pages that have methods other than GET and POST and also contain "crawler" word were also deleted.

- *Pageview Identification*

The pageview is identified using categorization based on the URL web page. The access log lines with the

"URL" field that contains "WP-admin" were deleted. The category code field based on URL was added to the database. Several category codes are described in Table 1.

Table 1. Page Category Code in Database

Category Code	URL	Description
4	/modul-pembuatan-aplikasi-berbasis-web.../	Tutorial about how to create web based application
5	/solusi-it/	IT solution that offered by LKP XYZ
10	/kelas-ramadhan-2016/	ICT training that held in Ramadhan 2016
12	/workshop-bulan-ini/	ICT training that held in August 2016
15	/kursus-komputer/	ICT course training list
23	/hubungi-kami/	Form to contact LKP XYZ
28	/category/program/	ICT Training program list
40	/cybersecurity/	One day seminar about cyber security
57	/tutorial-microsoft-excel/	Tutorial on absolute references, semi absolute dan relative in microsoft excel's formula
167	/how-do-i-configure.../	How to configure point to point bridge mode

- *Session Identification*

The access log data are sorted by visitor's date, time, and IP. The field user_id was added in order to count duration of session using Java programming. If a visitor access has duration less than 2 seconds, then the access log is deleted.

- *Data Transformation*

The user_id, session_id, and category code fields are transformed into the input format of SPMF tool. SPMF is an open data mining tool that is used in this study for performing sequential pattern mining. The algorithm used is Co-Occurrence Map Sequential Pattern Mining using the Bitmap Representation (CM-SPAM) algorithm.

Data preprocessing steps were done utilizing Java Programming using Eclipse Integrated Development Environment (IDE). MySQL and HeidiSQL were used for the database management system.

IV. Pattern Discovery Phase

The pattern discovery phase aims to find the visitor pattern that can be analyzed to extract information from web access log of the LKP XYZ's website. In this phase, sequential pattern mining was done to find the patterns.

A. Sequential Pattern Mining

Sequential pattern mining using the CM-SPAM algorithm was performed to find the sequence patterns of LKP XYZ website's visitor. The minimum support used was determined until the longest sequences were identified. The minimum support values were selected from 0.1% to 5%. The minimum support ratio and the sequence patterns resulted are shown in Table 2. The longest sequences are found at the minimum support of 0.01%, which results 7 sequence levels. The access log data in August 2016 has 27,133 visitors. At the minimum support of 0.01%, a subsequence is considered to be frequent of there are at least 27 records in the sequence dataset contain this subsequence.

1-Frequent Sequences

This study results 728 1-frequent sequences, where 15 frequent sequences have highest support value as shown in Table 3. The web page "/cybersecurity/" has the highest number of visitors in August 2016.

Based on the 1-frequent sequences, LKP XYZ website's visitors can be divided into 3 types. The first type is visitors who visit the "One-day seminar" and workshop page, the second type is visitors who access the tutorial / article page, and the third type is who visit other courses and training programs which are provided by LKP XYZ.

2-Frequent Sequences

Sequential pattern mining on the access log dataset results 293 2-frequent sequences. Five frequent sequences with highest support values is given in Table 4. The page "/cybersecurity/" is frequently visited after the page "/contact-us/". A new page appears on the 2-frequent sequences namely the page "/workshop-december-end-year-2013".

3-Frequent Sequences

This study obtains 137 3-frequent sequences. Five frequent sequences with highest support values are given in Table 5. The page "/cybersecurity/" was frequently visited after the page "/contact-us/". A new page appears on the 3-frequent-sequences namely the page "/category/program/".

4-Frequent Sequences and 5-Frequent Sequences

In addition to 3-frequent Sequences, as many 17 4-frequent sequence were found. Only one 4-frequent sequence meets the minimum support of 0.1%. This study also found 5 5-frequent sequences, and only one 5-frequent sequence meets the minimum support of 0.1%. In 4-frequent sequences and 5-frequent sequences, there are 3 most visited website pages namely "/contact-us/", "/cybersecurity/" and "/category/program/ ". Table 6 lists 4-frequent sequences and 5-frequent sequences generated at the minimum support of 0.01%.

Table 2. Frequent Sequences Generated at Minimum Support from 0.1% to 5%

No	*Minimum support* (%)	*Frequent Sequence*	*Frequent sequence number* (k)						
			1	2	3	4	5	6	7
1	0.10	728	272	293	137	17	5	3	1
2	0.20	185	112	60	8	3	2	0	0
3	0.30	116	79	31	3	2	1	0	0
4	0.40	75	54	17	2	1	0	0	0
5	0.50	53	39	11	2	1	0	0	0
6	0.60	43	32	7	2	1	0	0	0
7	0.70	31	22	6	2	1	0	0	0
8	0.80	27	19	6	2	0	0	0	0
9	0.90	24	18	5	1	0	0	0	0
10	1.00	22	17	4	1	0	0	0	0
11	1.10	22	17	4	1	0	0	0	0
12	1.20	20	16	3	1	0	0	0	0
13	1.30	17	16	3	1	0	0	0	0
14	1.40	16	12	3	1	0	0	0	0
15	1.50	14	12	3	1	0	0	0	0
16	1.60	14	11	2	1	0	0	0	0
17	1.70	11	9	2	0	0	0	0	0
18	1.80	10	8	2	0	0	0	0	0
19	1.90	9	7	2	0	0	0	0	0
20	2.00	9	7	2	0	0	0	0	0
21	2.50	6	4	2	0	0	0	0	0
22	3.00	6	4	2	0	0	0	0	0
23	3.50	5	4	1	0	0	0	0	0
24	4.00	5	4	1	0	0	0	0	0
25	4.50	4	4	0	0	0	0	0	0
26	5.00	4	4	0	0	0	0	0	0

B. Pattern Analysis

The recommendations proposed in this study consist of two parts, the first is recommendation based on sequential pattern mining and the second is general recommendation. The recommendations are proposed for LKP development and website development.

Based on sequential patterns from log access files, recommendations proposed to the LKP XYZ management are as follows:

- From the frequent sequences generated as shown in Table 2, the LKP XYZ management should improve their website to attract more visitors to explore the LKP XYZ website especially to promote the ICT training.
- The most visited page after "/contact-us/" is "/cybersecurity/". Therefore this study recommends the marketing team to held one day seminar periodically especially in the school holiday period.
- LKP XYZ management should promote the training schedule and materials on social media especially Facebook since 24% of the page "/cybersecurity/" visitors come from Facebook.

In addition to the LKP XYZ management, this study also proposes recommendations for the website developer. The recommendations are as follows

- The web page that advertises one day seminar or workshop through social media should also provide information about other training or courses material especially the short training or courses. Thus, the prospective learners can find easily information about seminar or workshop according to their preferences.
- Module or tutorial pages are also widely accessed by visitors of LKP XYZ website. Before visitors enter a certain module or tutorial, the website should ask visitors to provide their identity, so the visitors who download the specific module or tutorial can be informed about the related seminar or workshops. It also can be used for gauge the visitor trend. If the visitor trend to a certain module increases, then the LKP XYZ can provide training or seminar or workshop related to that module.

Based on sequence patterns of web visitors, this study proposes the following general recommendations for course management:

- Participants of one day seminar are awarded discount for the next seminar or workshop and for 1-year professional course or program organized

by the LKP XYZ. If the participant is a learner or a student from the LKP XYZ, he/she can also be awarded a special discount to follow the seminar or workshop.

- LKP XYZ website uses the Wordpress template. The recommendation for the website is to change the still image into gif type image, slide show, or video to attractively display the information.
- The naming structure of LKP XYZ website pages uses the alias. The website developer should choose an alias that more represents the content so web usage mining can provide more optimal results.

Table 3. 1-Frequent Sequences with Minimum Support of 0.01%

Sequence	Support (%)	Absolute Support
< {/cybersecurity/ }>	12.57	3410
< {/contact-us/ }>	4.46	1209
< {/module-developing-web-based-application.../ }>	2.90	787
< {/computer-course/ }>	2.54	690
< {/tutorial-microsoft-excel/ }>	0.84	227
< {/workshop-this-month/ }>	0.79	215
< {/page/2/ }>	0.78	212
< {/category/program/ }>	0.74	201
< {/course/ }>	0.70	189
< {/workshop-december-end-year-2013/ }>	0.64	173
< {/one-day-seminar-introduction-to-cybersecurity/ }>	0.63	170
< {/what-is-cloud-computing/ }>	0.57	155
< {/ramadhan-class-2016/ }>	0.52	141
< {/1-year-certification-program-it-internasional-3.. }>	0.48	131
< {/field-work-high-school/ }>	0.48	131

Table 4. 2-Frequent Sequences with Minimum support of 0.01%

Sequence	Support (%)	Absolute Support
< {/contact-us/}, {/contact-us/ }>	1.63	443
< {/cybersecurity/}, {/cybersecurity/ }>	1.18	319
< {/module-developing-web-based-application.../}, {/workshop-december-end-year-2013/ }>	0.55	149
<{/module-developing-web-based-application.../}, {/module-developing-web-based-application.../}>	0.45	123
< {/computer-course/}, {/computer-course/ }>	0.37	100

Table 5. Data 3-Frequent Sequence with Minimum Support 0.01%

Sequence	Support (%)	Support Absolut
< {/contact-us/}, {/contact-us/ }, {/contact-us/ }>	0.65	176
< {/cybersecurity/}, {/cybersecurity/ }, {/cybersecurity/ }>	0.32	88
< {/category/program/}, {/category/ program/}, {/category/program/ }>	0.15	42
<{/module-developing-web-based-application.../}, {/module-developing-web-based-application.../}, {/module-developing-web-based-application.../}>	0.10	28
< {/computer-course/}, {/computer-course/ }, {/computer-course/ }>	0.10	27

Table 6. 4-Frequent Sequences and 5-Frequent Sequences at Minimum Support of 0.01%

Sequence	Support (%)	Absolute Support
< {/contact-us/}, {/contact-us/ }, {/contact-us/ }, {/contact-us/ }>	0.28	76
< {/cybersecurity/}, {/cybersecurity/ }, {/cybersecurity/ }, {/cybersecurity/ }>	0.17	46
< {/category/program/}, {/category/ program/}, {/category/program/ }, {/category/program/ }>	0.10	28
< {/contact-us/ }, {/contact-us/ }, {/contact-us/ }, {/contact-us/ }, {/contact-us/ }>	0.12	32

V. Conclusions

This study has successfully implemented sequential pattern mining on LKP XYZ website access log to generate visitor patterns. The frequent sequence patterns are used to formulate the recommendations for the development of LKP XYZ managerial unit. However, with the current website structure, especially on the courses page, training materials that were accessed by the visitor cannot be extracted from the access log detail. A further research using web structure mining and web content mining should be done to obtain comprehensive recommendations for the LKP XYZ website improvement.

References

[1] B.M. Monjurul Alom, and M. Courtney, "Educational Data Mining: A Case Study Perspectives from Primary to University Education in Australia", *International Journal of Information Technology and Computer Science (IJITCS)*, Vol.10, No.2, pp.1-9, 2018.

[2] S.K Mohamad, and Z Tasir, "Educational data mining: A review". *The 9th International Conference on Cognitive Science. Procedia - Social and Behavioral Sciences*, vol. 97, pp.320 – 324, 2013.

[3] A. Soriano-Asensi, J. D. Mart ń-Guerrero, E. Soria-Olivas, A. Palomares. R. Magdalena-Benedito, and A. J. Serrano-L ópeza, "Web mining based on Growing Hierarchical Self-Organizing Maps: Analysis of a real citizen web portal", Expert Systems with Applications vol. 34, pp. 2988–2994, 2008.

[4] Z. Zhu, and J. Y. Wang, *Book recommendation service by improved association rule mining algorithm*, Proceedings of the sixth International Conference on Machine Learning and Cybernetics. Hong Kong (CN): Institute of Electrical and Electronics Engineers, pp. 3864-3869, 2007.

[5] J. Li, and P. Chen, *The application of association rule in library system*, International Symposium on Knowledge Acquisition and Modeling Workshop. Wuhan (CN): Institute of Electrical and Electronics Engineers, pp. 248-251, 2008.

[6] P. Jomsri, *Book recommendation system for digital library based on user profiles by using association rule*, Innovative Computing Technology, UK: Institute of Electrical and Electronics Engineers, pp. 130-134, 2014.

[7] I.S. Sitanggang, A. Agustina, N. A. Husin, and N. Mahmoodian, "Sequential Pattern Mining on Library Transaction Data," A paper presented in International Symposium on Information Technology 2010 (ITSim 2010), Kuala Lumpur, 15 - 17 June 2010.

[8] E. Garc á, C. Romero, S. Ventura, and C. de Castro, "A collaborative educational association rule mining tool," *Internet and Higher Education*, vol.14, pp.77–88, 2011.

[9] D. J. Ratnaningsih, and I. S. Sitanggang, "Comparative analysis of classification methods in determining non-active student characteristics in Indonesia Open University," *Journal of Applied Statistics*, vol. 43, no.1, pp. 87-97, 2016.

[10] J.V. Joshua, O.D. Alao, A.O.Adebayo, G.A. Onanuga, E.O. Ehinlafa, and O. E. Ajayi, "Data Mining: A Book Recommender System Using Frequent Pattern Algorithm," *Journal of Software Engineering and Simulation*, vol. 3, no. 3, pp. 01-13, 2016.

[11] L.M. Erman, and I.S. Sitanggang, "Clustering Undergraduate Computer Science Student Final Project Based on Frequent Itemset," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 8, no. 11, pp. 1-7, 2016.

[12] Y. Hung, K. B.Chen, C. Yang, and G. Deng, "Web usage mining for analysing elder self-care behavior patterns". *Expert Systems with Applications*, vol. 40, pp. 775–783, 2013.

[13] P.F. Viger, G. Antonio, M. Campos, and R. Thomas, *Fast Vertical Mining of Sequential Patterns Using Co-occurrence Information*, PAKDD part 1. Switzerland: Springer International Publishing, pp. 40-52, 2014.

[14] R. Kang, J. Radinsky, and L. Lyons, "Frequent Sequential Interactions as Opportunities to Engage in Temporal Reasoning with an Online GIS". *ACM* 978-1-4503-3417-4/15/03, 2015.

[15] S. Djamasbi, T. Tullis, M. Siegel, F. Ng, D. Capozzo, and R. Groezinger, "Generation Y & Web Design: Usability Testing through Eye Tracking". Proceedings of the Fourteenth Americas Conference on Information Systems (AMCIS), Toronto, Canada, pp.1-11, 2008.

Authors' Profiles

Susi Maulidiah is a master student in computer Science, Bogor Agricultural University. Her research interest is on data mining.

Imas Sukaesih Sitanggang is a lecturer in Computer Science Department, Bogor Agricultural University, Indonesia. Her main research interests include spatial data mining and data warehousing.

Heru Sukoco is a lecturer in Computer Science Department, Bogor Agricultural University, Indonesia. His main research interests include future internet, network infrastructure performance and evaluation, embedded and control system, and wireless sensor networks

How to cite this paper: Susi Maulidiah, Imas S. Sitanggang, Heru Sukoco, "ICT Training Recommendation using Web Usage Mining", International Journal of Information Technology and Computer Science(IJITCS), Vol.10, No.12, pp.21-26, 2018. DOI: 10.5815/ijitcs.2018.12.03