Research Article

International Journal of Advanced Computer Research, Vol 6(27)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2016.627009

Designing web-based data mining applications to analyze the association rules tracer study at university using a FOLD-growth method

Herman Yuliansyah* **and Lisna Zahrotun**
Department of Informatics, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

Received: 22-July-2016; Revised: 26-September-2016; Accepted: 28-September-2016

Abstract
Tracer study is one of strategy made by university to obtain information and feedback from alumni and stakeholders to measure educational outcomes and improve the process of education and learning in the future. Data mining can be used as a way to collect data feedback from alumni and stakeholders to obtain some useful information. This study has used fast online dynamic-growth (FOLD-growth) to find correlations between different data feedback to determine the pattern of association. The methodology in this study refers to software development methods: analysis system, design system, implementation system and testing system. The result of this study is a design of user interface for web-based data mining applications to analyze the association rules tracer study at university using FOLD-growth method and the design has been tested the acceptance using the method system usability scale (SUS).

Keywords
Data mining, Association rules, Tracer study, FOLD-growth method.

1.Introduction
Tracer study is one of strategy made by university to obtain information of graduates, so it can evaluate the educational process, measuring the educational goals and make an improvement in the future and in order to establish where graduates are, do, and what interventions can be made to improve their professional activities and services [1-3]. Meanwhile, data mining is a process to find patterns and trends conducted with pattern recognition technology, statistical and mathematical techniques to sort out a number of useful data in a data set [4, 5]. Association rule is a pattern of association among itemsets, it is a fundamental task and is of great importance in many data mining applications [6]. One of the algorithms that can be used to determine the rule is fast online dynamic-growth (FOLD-growth). It is a combination of fast online dynamic association rule mining (FOLDARM) and frequent pattern (FP)-growth. Several studies in data mining have been carried out. Babu et al. [7] proposed a tree that is fast and efficient in frequent pattern extraction.

This paper shows the advantages of FP-Growth algorithm for association rule mining by using the newly proposed approach [7]. Gao et al. [8] took the historical data of university graduate employment and tried to find out the useful information hidden through the data mining. The results in this study were the preliminary study of the historical data on employment and established the subject. The data mining model is then built to do the mining analysis by using data mining tool name Weka [8].

This paper focused on designing web-based data mining applications which are capable to analyze the association rules of tracer study using FOLD-growth method. The data, which have been used, can be related to how long to get a job, study duration, age, skill, grade point average and the first salary.

2.Materials and methods
2.1Materials
2.1.1Association rule
One measure of performance for an association rule "A => B" is the amount of support that is denoted by (A => B) which can be interpreted as "If A, then B". "A" is the antecedent (the predecessor of the implications), while B is called the consequent (followers of implications). The calculation of the

*Author for correspondence

This research has been supported by Lembaga Penelitian dan Pengembangan Universitas Ahmad Dahlan in the scheme of Penelitian Kerjasama Kelembagaan (PKK) number: PKK-037/LPP-UAD/III/2016.

amount of support is defined according to equation 1 [9].

$$s(A => B) = P(A \cap B) = \frac{The\ number\ of\ transactions\ containing\ items\ on\ A \cap B}{The\ total\ number\ of\ transactions} \quad (1)$$

Other performance measures for association rules "A => B" is the amount of support that is denoted by conf (A => B) and is defined according to the equation 2.

$$conf(A => B) = P(A|B) = \frac{The\ number\ of\ transactions\ containing\ items\ on\ A \cap B}{The\ number\ of\ transactions\ containing\ items\ in\ A} \quad (2)$$

An item set is a set that consists of some or all of the items that are members. An itemset consisting of k items is called k-itemset.

2.1.2FOLDRAM

FOLDRAM is a data-mining algorithm that uses a data structure support-ordered Trie item-set (SOTrieIT) and it allows generating large 1-itemsets and 2-itemsets quickly without scanning the database [10].

2.1.3FP-tree

FP-tree is a data structure that is built by mapping each data transaction on any specific track, so every transaction that has the same items will be compressed to overwrite each other. An advantage of FP-Tree is it only scanning the data transaction twice.

2.1.4FOLD-growth

FOLD-growth algorithm is combination of FOLDARM and FP-growth algorithm. FOLDARM has fast performance when size item set frequent maximum (kmax) is small or kmax = 10 with and FP-Growth which has a fast performance when kmax > 10. FOLD-growth algorithm has four main stages [11]:

- Mining L1 and L2 using SOTrieIT
- Pruning items that are not frequent
- Build FP-tree using puned transactions
- Mining frequent item set with FP-growth algorithm.

2.2Methods

The methodology referring to software development life cycle:

2.2.1Requirement analysis

The results of this phase are the functional and non-functional requirements from the user.

2.2.2Design software

This phase will be the user interface, which connects the user to the system.

2.2.3Software implementation

This phase involves in creating the software by write the source code.

2.2.4Software testing

In this phase the software has been tested using the system usability scale (SUS) method. It is used as self-administered instrument for the evaluation of the usability in wide range of products and user interfaces [12].

3.Results

3.1Data set

Table 1 is a sample of the dataset. *Table 1* shows the alumni database consist of;

A: represent a time to get a job
B: represent the period of study
C: represent age
D: represents the skills of English
E: represents the skill of fields
F: represents a grade point average (GPA)
G: represents first salary work

Identity of A, B, C, D, E, F, and G are described in *Table 2*.

Table 1 Examples of alumni data

Student number	Alumni data
0900210	A1, B2, C2, D1, E3, F2, G2
0900211	A3, B2, E3, F2
0900212	A2, B2, D2, E3, F2, G1
0900213	A3, B1, C2, F2, G3
0900214	A1, B2, C3, E1, F1, G2
0900215	A2, B3, C1, D1, E3, F1, G1
0900216	A1, B2, C2,D2, F2
0900217	A3, B1, C1, D1, F2, G3
0900218	A1, B2, C2, F2, G3
0900219	A1, B3, C2, D1, E3, F2, G2

Table 2 Reference value of alumni data

Time to get a job		Study duration		Age		English skill		Field skill		Grade point average		First salary	
<=6 Month	**A1**	3-4 Year	**B1**	17-20 Year	**C1**	Excellent	**D1**	Excellent	**E1**	**> 3,50**	**F1**	<1 Million	**G1**
6-12 Month	**A2**	4-5 Year	**B2**	20-23 Year	**C2**	Good	**D2**	Good	**E2**	**2,75 - 3,50**	**F2**	1-5 Million	**G2**
>12 Month	**A3**	5-7 Year	**B3**	>23 Year	**C3**	Fair	**D3**	Fair	**E3**	**< 2,75**	**F3**	>5 Million	**G3**

3.2FOLD-growth

There are four phases in the determination of the association rule FOLD-growth:

3.2.1Mining L1 and L2 using SOTrieIT

In the first phase, the database has been scanned. For each data, L1 (1-itemset) and L2 (2-item-set) is generated. It is recorded in SOTrieIT. L1 is the first item data from the data sets, which calculated by the appearance or called by the count (number). The results shown in *Table 3* is the amount of each item set, while L2 is a combination of two items from the list L1 which meets the minimum count = 4. *Table 4* shows the results of the amount combination of two items.

Table 3 List of L1 (1-itemset)

Alumni data	Count
A1	5
A2	2
A3	3
B1	2
B2	6
B3	2
C1	2
C2	5
C3	1
D1	4
D2	2
D3	0
E1	1
E2	0
E3	5
F1	2
F2	4
F3	0
G1	2
G2	3
G3	3

Table 4 List of L2 (2-itemset)

Item name	Count
A1, B2	4
A1, C2	5
A1, D1	4
A1, E3	5
A1, F2	7
B2, C2	3
B2, D1	1
B2, E3	3
B2, F2	3
C2, D1	2
C2, E3	1
C2, F2	5
D1, E3	3
D1, F2	3
E3, F2	4

3.2.2Pruning the not frequent items

In the second phase, the not frequent items from the database are pruned. Each data, which has more than 2, will be checked using the L1 and L2 that had been stored in SOTrieIT. Thus, the items that are not frequent are terminated. In this case the minimum limit is 4. Pruning data using L1 and L2 are called the frequent ordered items. *Table 5* shows frequent ordered list items that have been sequenced.

Table 5 List of ordered frequent items

Student number	Alumni data
0900213	C2, F2
0900216	A1, F2
0900210	A1, F2, E3
0900211	F2
0900212	F2
0900214	A1, B2
0900218	A1, F2, C2
0900219	A1, F2, C2

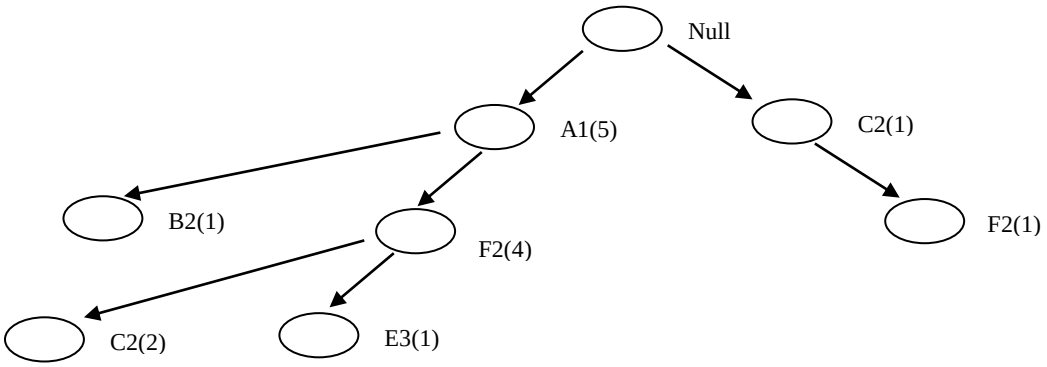

Figure 1 Results of FP-tree frequent items ordered

3.2.3Build FP-tree using the data that has been pruned

In the third phase, FP-tree will be built using the data that have been pruned or frequent ordered items in *Table 5*. *Figure 1* is a form of FP-tree obtained from *Table 5*.

3.2.4 Mining frequent itemset with FP-growth algorithm

In the fourth phase, data will be calculated using the conditional pattern based and the generation of conditional FP-trees. Conditional pattern based is sub database, which contain the final track and suffix patterns. *Table 6* shows the results of the conditional based pattern, which support the count of each item data in conditional pattern based. Data will be selected to meet the equal support count number value or more than the minimum support count. Then, in *Table 7* each selected item set will be generated with conditional FP-Tree.

Table 6 Conditional pattern based

Item name	Conditional pattern based	Count
C2	F2, A1	2
B2	A1	1
E3	F2, A1	1
F2	C2	1

Table 7 Conditional FP-tree

Item name	Conditional FP-tree	Count
C2	F2	2
C2	A1	2
C2	F2, A1	2

Table 8 Frequent itemset

Item name	Frequent itemset	Count
C2	F2	2
C2	A1	2
C2	F2, A1	2

The search of the frequent itemset conducted with the combination of items for every single track of the conditional FP-tree shown in *Table 8*. After the frequent item set is obtained, the next step is to calculate the value of the support and confidence that the candidate of the association rules.

Support values have been calculated using equation 1 and the value of confidence will be calculated using the equation 2 and the results can be seen in *Table 9*. The result of the prospective association rules is shown in *Table 8*.

From *Table 9*, the rules that have a value less than the minimum support and confident will be pruned. In this study, to obtain the association rules as shown in *Table 10* than the minimum support value should be 25% and confident should be 60%.

Table 9 List of candidate association rules

From frequent itemset	Association rules	Support		Confidence	
C2, F2	If C2 Then F2	2/8	25%	2/3	67%
	If F2 Then C2	2/8	25%	2/5	20%
C2, A1	If C2 Then A1	2/8	25%	2/3	67%
	If A1 Then C2	2/8	25%	2/5	20%
C2, F2, A1	If C2 And F2 Then A1	2/8	25%	2/3	67%
	If C2 And A1 Then F2	2/8	25%	2/2	100%
	If F2 And A1 Then C2	2/8	25%	2/3	67%

Table 10 List of candidate association rules

From frequent itemset	Association rules	Support		Confidence	
C2, F2	If C2 Then F2	2/8	25%	2/3	67%
C2, A1	If C2 Then A1	2/8	25%	2/3	67%
C2, F2, A1	If C2 And F2 Then A1	2/8	25%	2/3	67%
	If C2 And A1 Then F2	2/8	25%	2/2	100%
	If F2 And A1 Then C2	2/8	25%	2/3	67%

4.Discussion

Based on the results of the university's tracer study analysis using association rules and FOLD-growth method, the design of user interface for the application can be made. *Figure 2* shows the data of all the alumni that will be used as a data set in the data mining process. This data set obtained from the alumni information system. This data set is the result of a survey process of the university's alumni through the alumni information system. The data shown in *Figure 2* will be processed into a 1-itemsets and 2-itemsets. This process refers to *Table 2* and *Table 3* so that the results can be displayed as shown in *Figure 3*. *Figure 3* is the result of the calculation of the emergence of the data set and will display the data items that meet the minimum number count is 4.

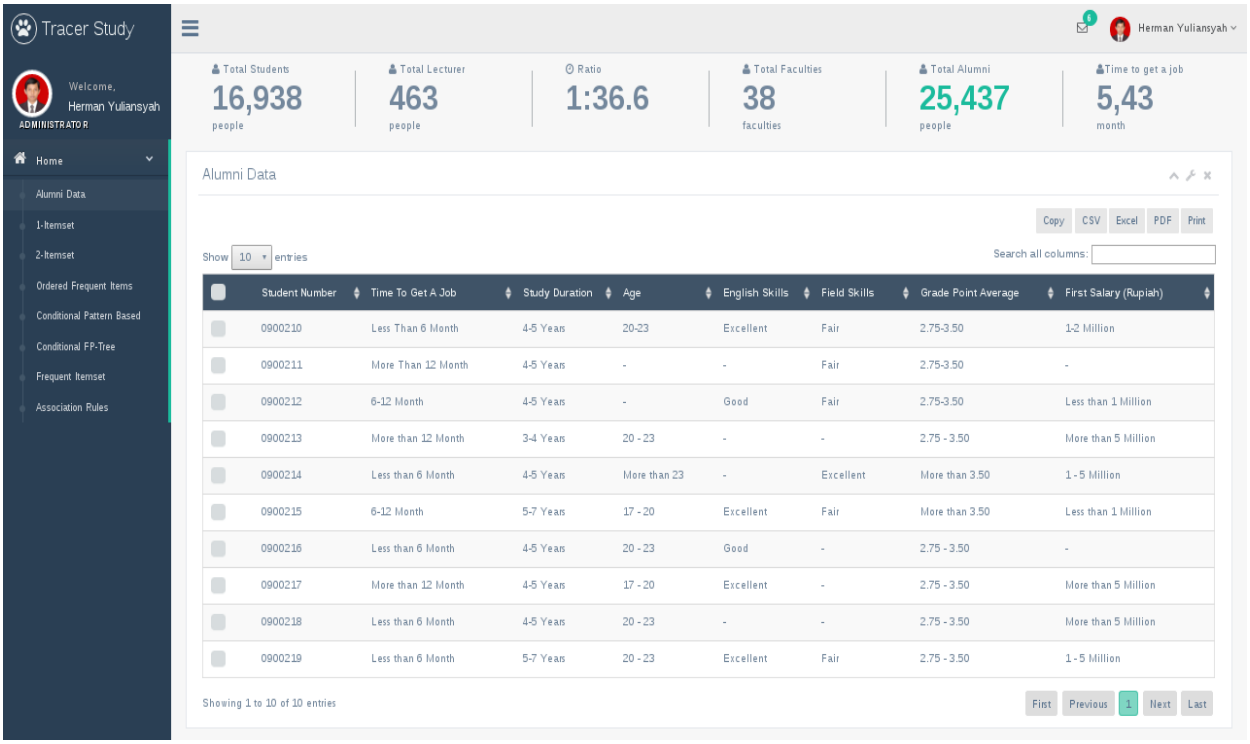

Figure 2 Tabulation alumni data

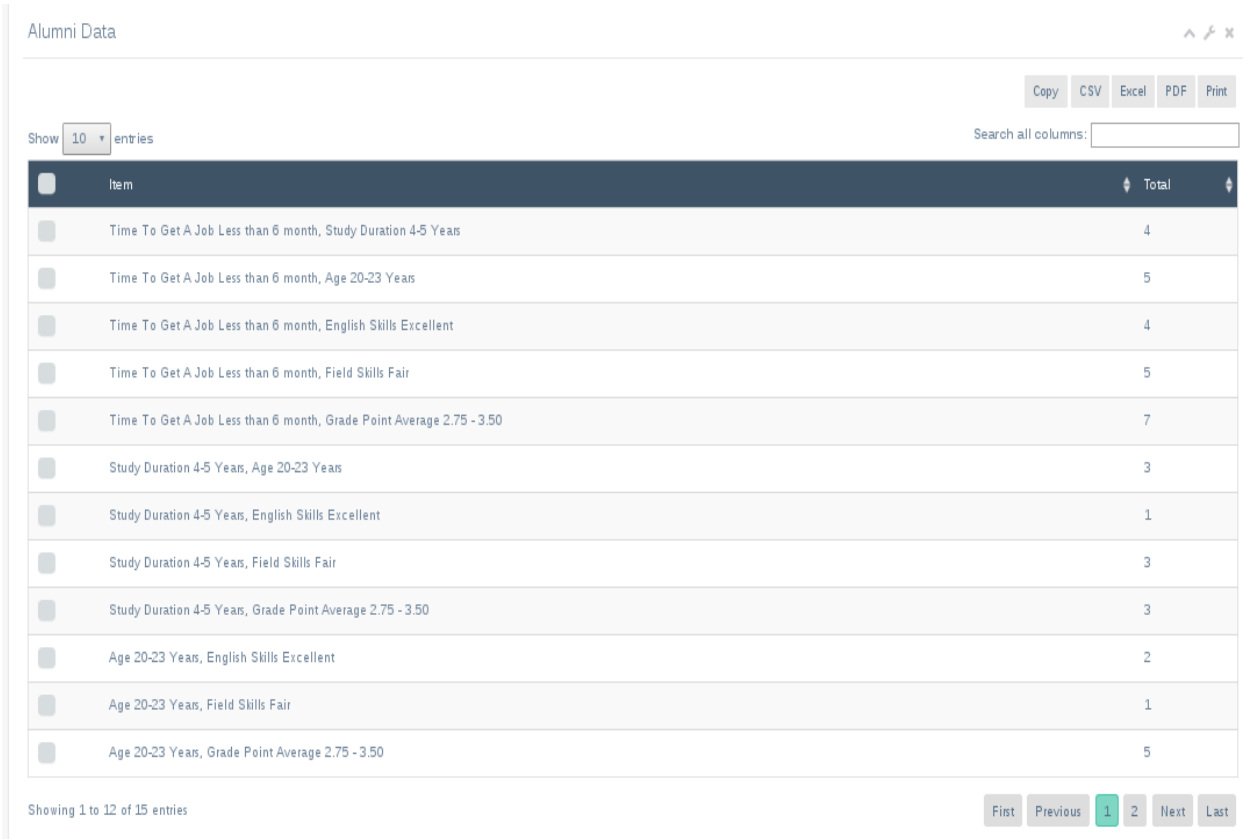

Figure 3 Tabulation L2 (2-itemset)

Figure 4 Tabulation frequent itemset

The process of data mining followed by frequent item set search process with FP-growth algorithm is shown in *Figure 3*. *Figure 4* obtained from the generation process conditional pattern-based and conditional FP-tree. *Figure 5* displays the results of the association rules candidate that meet minimal support and confidence values. *Figure 5* was obtained from the frequent item set where the support and confidence value of the data has been calculated as a candidate for association rules. If the support and confidence value has less value than the minimal support and confidence should, then the data will be trimmed. *Table 11* shows a user acceptance test for user interface design with standard SUS. The standard SUS consists of the following ten items. To use the SUS, present the items to participants as 5-point scales, numbered from 1 (anchored with "Strongly disagree") to 5 (anchored with "Strongly agree"). For positively worded items (1, 3, 5, 7 and 9), the score contribution is the scale position minus 1. For negatively worded items (2, 4, 6, 8 and 10), it is 5 minus the scale position [13] [14].

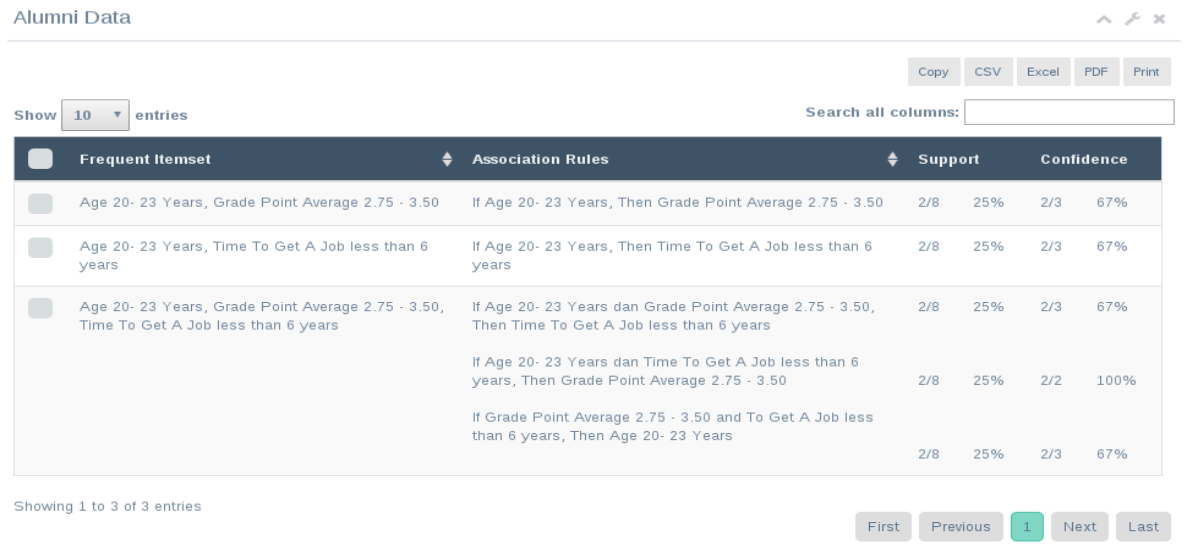

Figure 5 Tabulation candidate association rules

5.Conclusion

The result of this study is a design of user interface for web-based data mining applications to analyze the association rules tracer study at university using FOLD-growth method. The result of this design has been tested the acceptance using the method SUS. So that the design of these applications can be accepted and implemented as a web-based data mining applications which integrated with alumni information systems as its feeder dataset.

Table 11 System usability scale

No	Statement	Strong disagree				Strong agree
1	I think that I would like to use this system frequently	1	2	3	4	5
2	I found the system unnecessarily complex	1	2	3	4	5
3	I thought the system was easy to use	1	2	3	4	5
4	I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5	I found the various functions in this system were well integrated	1	2	3	4	5
6	I thought there was too much inconsistency in this system	1	2	3	4	5
7	I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
8	I found the system very cumbersome to use	1	2	3	4	5
9	I felt very confident using the system	1	2	3	4	5
10	I needed to learn a lot of things before I could get going with this system	1	2	3	4	5

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments and suggestions that helped to improve the quality and presentation of this paper.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Chandra R, Ruhama S, Sarjono MW. Exploring tracer study service in career center web site of Indonesia higher education. International Journal of Computer Science and Information Security. 2013;11(3):36-9.

[2] Shongwe M, Ocholla D. A tracer study of LIS graduates at the University of Zululand. In the 6th biennial ProLISSA conference 2011.

[3] Abidin M. Alumni satisfaction on curriculum structure and learning process in Indonesian Islamic University. International Journal of Scientific Research and Education. 2015;3(2):2900-5.

[4] Gartner Group. http://www.gartner.com/it-glossary/data-mining/. Accessed 26 May 2016.

[5] Larose DT. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons; 2014.

[6] Sahoo J, Das AK, Goswami A. An effective association rule mining scheme using a new generic basis. Knowledge and Information Systems. 2015; 43(1):127-56.

[7] Babu DB, Prasad RS, Umamaheswararao Y. Efficient frequent pattern tree construction. International Journal of Advanced Computer Research. 2014;4(1):331-6.

[8] Gao L. Analysis of employment data mining for university student based on WEKA platform. Journal of Applied Science and Engineering Innovation. 2015;2(4):130-3.

[9] Susanto S, D. Introduction to data mining gain knowledge of a chunk of data. CV ANDI OFFSET. Yogyakarta. 2010.

[10] Woon YK, Ng WK, Das A. Fast online dynamic association rule mining. In proceedings of the second international conference on web information systems engineering 2001 (pp. 278-87). IEEE.

[11] Soelaiman R, WP NM. Analysis algorithm performance-fold growth and FP-growth on an excavation pattern association. National seminar in application of information technology (SNATI), 2006.

[12] Martins AI, Rosa AF, Queirós A, Silva A, Rocha NP. European portuguese validation of the system usability scale (SUS). Procedia Computer Science. 2015:293-300.

[13] Brooke J. SUS-a quick and dirty usability scale. Usability evaluation in industry. 1996; 189(194):4-7.

[14] Lewis JR, Sauro J. The factor structure of the system usability scale. In international conference on human centered design 2009 (pp. 94-103). Springer Berlin Heidelberg.

Herman Yuliansyah is a Lecture of informatics department, Universitas Ahmad Dahlan, Yogyakarta, Indonesia. Herman graduated from Universitas Muhammadiyah Yogyakarta, Indonesia in 2007 and received a bachelor of electrical engineering degree. Then entered Universitas Gajah Mada, Indonesia and received a master of engineering degree in 2011. The major field of study is Software Engineering and Web Mobile Application Development.
Email: herman.yuliansyah@tif.uad.ac.id

Lisna Zahrotun is a Lecture of informatics Department, Universitas Ahmad Dahlan, Yogyakarta, Indonesia. Lisna graduated from Universitas Ahmad Dahlan, Indonesia in 2007 and received a bachelor of informatics degree. Then entered Universitas Gajah Mada, Indonesia and received a master of computer science degree in 2014. The major field of study is data mining.