Implementation of Association Rule in Repeated Courses

(Case Study: Telkom Applied Science School)

Tora Fahrudin
Telkom University, Telkom Applied Science School,
Department of Information Technology, Bandung
torafahrudin@telkomuniversity.ac.id

Abstrak*— Sebuah Perguruan tinggi seharusnya sudah mempunyai matakuliah yang tersusun kedalam sebuah kurikulum. Tiap matakuliah mempunyai karakteristiknya masing-masing, mulai dari mudah maupun yang susah. Hal tersebut terlihat dari rata rata ketidaklulusan tiap matakuliah. Di Fakultas Ilmu Terapan, rata rata prosentase ketidaklulusan tiap matakuliah seluruh semester bervariasi dari 0% sampai 59.13%. Akan tetapi belum ada studi di Fakultas, mengenai implikasi antara sebuah matakuliah yang diulang dengan matakuliah diulang yang lain. Salah satu cabang dari data mining adalah asosiasi, dimana tujuannya adalah menemukan ekspresi implikasi X->Y, dimana X dan Y adalah himpunan itemset yang saling bebas. Dengan mengimplementasikan teknik asosiasi menggunakan software KNIME, dapat terlihat kemunculan ekspresi implikasi yang memenuhi minimum support (s) dan confidence (c). Dari hasil penelitian dapat disimpulkan : 1) implementasi asosiasi di universitas adalah menemukan ekspresi implikasi antara sebuah matakuliah yang diulang dengan matakuliah yang diulang lainnya. 2) support dan value yang di temukan bervariasi dari 1%-6%. Sedangkan confidence 6%-9%. 3) jika setting nilai minimum support semakin kecil, maka akan banyak didapatkan rule asosiasi. 4) matakuliah statistic adalah paling banyak diulang bersamaan dengan praktikum statistic dengan nilai support 1% dan nilai confidence 94%.

Kata Kunci— Asosiasi, KNIME, Data Mining, Support, Confidence, matakuliah di ulang

***Abstract*— College must have had well organized course in its curriculum. Each course has its own characteristics, ranging from easy to difficult. The characteristics are seen from the various failure rates of each course. At Telkom Applied Science School, the average percentage of failure rate of each course of all semesters varied from 0% to 59.13%. However, there is no study in Telkom Applied Science School which examined the implication expression between the taking of one repeated course and other repeated courses. One branch of data mining is association that is the pattern of the implication expression of the form X->Y, where X and Y is disjoint item sets. By implementing association techniques using KNIME software, it can be seen the emergence of an item set and other item set that meet the minimum support (s) and confidence (c) thresholds. From this research, it can be concluded that: 1) one of association rule implementation on University is to find relationship between the taking of one repeated course and other repeated courses. 2) Support (s) value varies from 1% to 6%. While, confidence (c) value varies from 6% to 94%. 3) If the support value threshold is smaller, there will be more association rule and 4) Statistics is course mostly repeated by students together with statistics practice with support value (s) 1% and confidence value (c) 94%.**

Keywords*— *Association, KNIME, Data Mining, Support, Confidence, repeated a course

I. INTRODUCTION

College must have had well organized course in its curriculum. Each course has its own characteristics, ranging from easy to difficult. The characteristics are seen from the various failure rates of each course.

Telkom Applied Science School, which concern on IT to make the students mastering branch of Science and or Technology to meet the national interest and increase the nation's competitiveness [1], must have many lecturers who have IT competencies such as programming. However, programming is not an easy subject to be studied. It requires correct understanding of abstract concepts [2]. Also the students are heterogeneous and thus it is difficult to design the instruction. This often leads to high failures rates in programming courses [2].

In Telkom Applied Science School of Telkom University, the average percentage of failure rate of each course of all semesters varied from 0% to 59.13% [3]. The following table shows the average percentage of number of students who failed in each course of all semesters [3]. Repeated courses means student who failed in one course, must take that course again in other semester.

From Table I, it can be seen the list of courses that has high failure percentages. The information on Table I can be easily generated from the database by doing query. However, query can’t be used to see the implication expression between the taking of one repeated course and other repeated courses. Thus, association technique of data mining is used in this research in order to provide solutions for the problem above.

TABLE I. AVERAGE PERCENTAGE OF NUMBER OF STUDENTS WHO FAILED IN EACH COURSE OF ALL SEMESTERS

Subject Code	Percentage
MI3504 (Mobile Technology)	59.13
BC 123 (Religion: Protestant)	41.06
MI3234 (XML Data Processing)	39.23
MI2283 (Object Oriented Programming)	35.59
TK3323 (Rest Programming)	35.58
……….	….....
KA3242 (Advanced Financtial Accounting)	0
KA3293 (Advanced Tax)	0

II. ASSOSIATION DATA MINING

Association rule is one of some method from the Data mining concept [4]. Association is process to find relationship between one attribute to another attribute on record data [4]. In the association, some concepts of implication expression for X Y, where is X and Y an item set:

1. Frequent pattern: A pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set.
2. Item set X: set of item, such as $\{x1,x2,\ldots x_k\}$
3. Support (s): Probability that a transaction contains $X \cup Y$
4. Confidence (c) : Conditional probability that a transaction having X also contain Y
5. The goal from association is to find all rules X --> Y with satisfy minimum support (s) and Confidence (c). [4]

As shown in Table II, is example of sales transaction data for the market basket analysis.

TABLE II. TRANSACTION AND ITEM SET EXAMPLE [4]

Transaction Id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

TID is transaction ID, and item set is collection of item bought by customer. TID 10 buys 3 items, such as item A, C and item D.

Minimum support (s) = 50% (1)

Minimum confidence (c) = 50% (2)

Will be A-> D (support 60%, confidence 100%),

D-> A (support 60%, confidence 75%).

We can say that item A and item D will appear 3 times from all transaction id (TID), so support (s) is (3/5) X 100% = 60%. From 3 times appearance of item A in all transaction, item D appear to all transaction, so confidence (c) is 100% (3/3 x 100%) [4].

By using the analogy from sales transaction data on Table II, the same concept can be used to explore the association rule from the taking of one repeated course and other repeated courses. Transaction ID can be represented by student number. Meanwhile, item set can be represented by list of courses which repeated by student.

TABLE III. LIST OF REPEATED COURSES OF EACH STUDENT [4]

Student Number	List of repeated courses
30107001	TE112, CA112, BC172, BC293
30107002	BC022, BC202, IS551, CE122, IS511, CA202, IS303, BC192, CA112
30107003	IS242, CA122
30107004	BC162
……………	………..

From the table III , it shows that student with student number 30107001 retake 4 courses: TE112, CA112, BC172 and BC293. So from that table, we can use association rule to find implication expression between the taking of one repeated course and other repeated courses. Until current decade, association techniques still evolve. The modifications made more effective techniques for various fields [8].

III. MINING ASSOCIATION RULE OF REPEATED COURSES

A. Data Preprocess (format the data to transaction item set)

The first step of the data mining method is preprocessed. The purpose of preprocessing is getting data ready for a machine learning algorithms. Association Rule requires the data to be represented to Table 2 format (transaction id and a list of transactions). Source of data was taken from the Student Study Card table that contains class, subject_code, semester, school_year and student_number fields.

TABLE IV. STUDENT STUDY CARD TABLE

Class	Subject Code	Semester	School Year	Student Number
AIS-08-03	IS252	4	2008-2009	30107001
AIP-0807	IS113	1	2008-2009	30107001
CIT-0804	BC293	5	2008-2009	30107001
EAP-09-01	IS571	4	2009-2010	30107001

Table IV shows that only repeated courses data (each student number and subject code more than 1 record) would be processed. By using GROUP BY and HAVING, each student number and each subject code could be grouped, counted and filtered.

TABLE V. REPEATED COURSES OF EACH STUDENT FROM STUDENT STUDY CARD TABLE

Student Number	Subject Code	Record Count
30107001	TE112	2
30107001	CA112	2
30107001	BC172	3
30107001	BC293	2
…..	……	….

Table V should have been formatted to table 3 before processed by association rule algorithm. The step by step transformation:

- Create table course_point_pre_process from the Table V and export to flat file format.

TABLE VI. COURSE POINT PRE PROCESS TABLE

Student Number	Subject Code
30107001	TE112
30107001	CA112
30107001	BC172
30107001	BC293
…..	……

- Make a group list of a course, each student number with Group By with Group By Column is Student number and Group by setting Subject course, on KNIME software (Fig 1 and 2) [5].
-

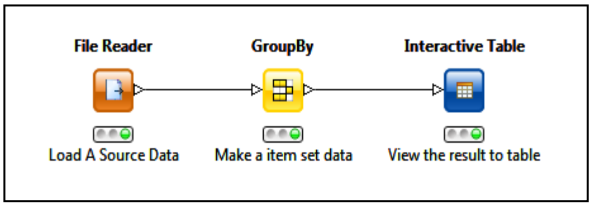


Fig. 1. Workflow on KNIME software to preprocess the flat file data into item set data

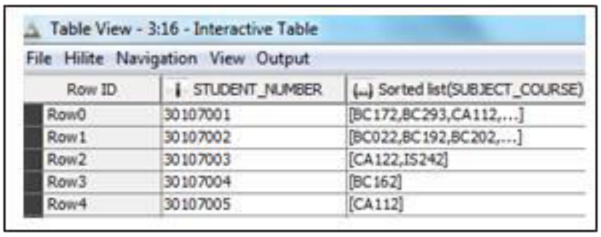


Table View - 3:16 - Interactive Table
File Hilite Navigation View Output

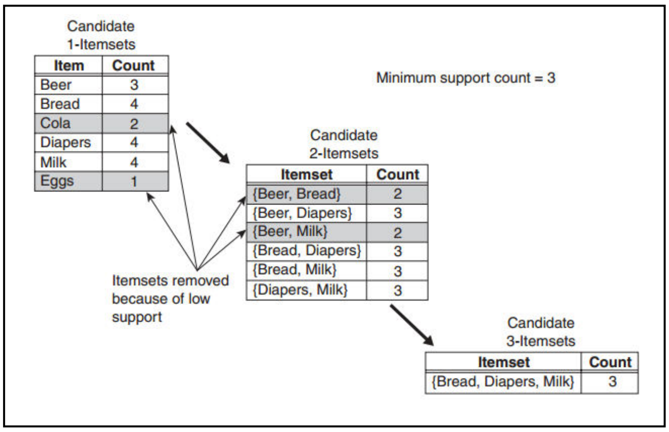
Row ID	STUDENT_NUMBER	Sorted list(SUBJECT_COURSE)
Row0	30107001	[BC172,BC293,CA112,...]
Row1	30107002	[BC022,BC192,BC202,...]
Row2	30107003	[CA122,IS242]
Row3	30107004	[BC162]
Row4	30107005	[CA112]

Fig. 2. Item set format

B. *Association Rule Learner (Apriori Algorithm)*

Association Rule Learner in KNIME use Apriori Algorithm [5]. Apriori is the first association rule mining algorithm that pioneered the use of support-based pruning to systematically control the exponential growth of candidate item sets [6].

As shown in Figure 3, is some example of apriori algorithm to generate frequent item sets for Market Base Transaction data


Minimum support count = 3
Itemsets removed because of low support

Candidate 1-Itemsets

Item	Count
Beer	3
Bread	4
Cola	2
Diapers	4
Milk	4
Eggs	1

Candidate 2-Itemsets

Itemset	Count
{Beer, Bread}	2
{Beer, Diapers}	3
{Beer, Milk}	2
{Bread, Diapers}	3
{Bread, Milk}	3
{Diapers, Milk}	3

Candidate 3-Itemsets

Itemset	Count
{Bread, Diapers, Milk}	3

Fig. 3. Illustration of frequent item set using Apriori Algorithm [6]

Important terms used in Apriori [9]:

• Min_sup (minimum support) : it is minimum support used for searching frequent patterns that satisfy this constraint

• Min_conf (minimum confidence) : it is minimum confidence used for finding the strong association rule that satisfy this threshold

One of weakness from apriori algorithm is not efficient for large dataset. In case of large dataset, Apriori Algorithm produce large number of candidate itemsets. Algorithm scan database repeatedly for searching frequent itemsets, so more time and resource are required [9]. Apriori algorithm still been developed for improving efficiency. Six techniques was introduced in bhandari survey paper [10]. In this research, apriori algorithm was used because the data not too large.

On KNIME Association rule learner, some input variable such as Minimum Support (s) and Minimum Confidence (c) needed to be set first before running. Minimal support (s) and confidence (c) is set to 1%. Min (s) = 0.01 and Min (c) = 0.01.

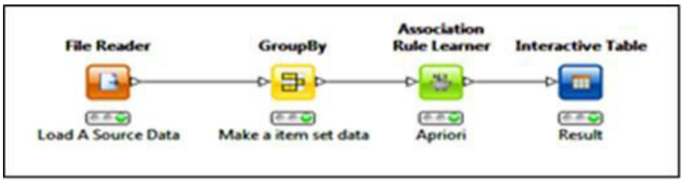


Fig. 4. Association Rule Learner Workflow

Node Association rule Learner is one of 12 Mining Node on KNIME. By using association rule learner node, association rule will be extracted from transactions which qualify the minimum support and confidence.

C. *Extracted rule*

The results obtained from association rule Learner KNIME are 608 rules, with value of support (s) varied from 1% to 6%.

TABLE VII. SUPPORT AND CONFIDENCE VALUE FROM EXTRACTED RULE

No	Support	Min Confidence	Max Confidence	Rule Count
1	0.01	0.06	0.94	366
2	0.02	0.09	0.66	190
3	0.03	0.15	0.69	38
4	0.04	0.19	0.47	10
5	0.05	0.37	0.38	2
6	0.06	0.32	0.44	2
			Total	**608**

Table VII shows that max support (s) is 0.06 (6%) from all repeated course transaction, while max confidence value is 0.94 (94%). It is seen that the smaller value of support (s), then more association rules would be obtained. The small value of support (s) indicates that the implications X—> Y only occurs in 1-6% from all transactions. Meanwhile greater value of confidence (c) make generated rule more interesting / promising. Related to this research purpose, our goal is to look for repeated course that have a great relationship with other repeated courses. Rules will considered if a rule has confidence (c) values more than 70%.

There are 41 associations rule are found with the rule confidence value >= 0.7 (equal or more than 70%).

TABLE VIII. THE RULE COUNT FOR CONFIDENCE (C) VALUE >= 0.7

Confidence Value	Rule Count
0.7	14
0.8	2
0.9	1
Total	**41**

Rule generated with confidence (c), value >= 0.7 can be seen in the following table (S = Support, C = Confidence)

TABLE IX. RULE GENERATED FOR CONFIDENCE VALUE >= 0.7 (WITH SUBJECT CODE)

No	(S)	(C)	Antecedent	Implies	Consequent
1	0.01	0.94	[MF511]	-->	MF142
2	0.01	0.8	[IS581]	-->	IS242
3	0.01	0.8	[CE531]	-->	CE143
4	0.01	0.7	[IS143 IS521]	-->	IS162
5	0.01	0.7	[MF113 IS521]	-->	MF133
6	0.01	0.7	[IS162 IS132]	-->	IS143
7	0.01	0.7	[MF113 IS541]	-->	IS182
8	0.01	0.7	[MF133 CE113]	-->	MF113
9	0.01	0.7	[MF133 BC162]	-->	MF113
10	0.03	0.7	[IS541]	-->	IS182
11	0.01	0.7	[IS162 BC132]	-->	MF133
12	0.01	0.7	[BC162 IS182]	-->	MF113
13	0.01	0.7	[IS113 BC113]	-->	MF113
14	0.01	0.7	[MF133 IS541]	-->	IS182
15	0.01	0.7	[IS143 IS521]	-->	MF133
16	0.02	0.7	[MF133 IS521]	-->	IS162
17	0.01	0.7	[IS162 BC132]	-->	IS143

TABLE X. RULE GENERATED FOR CONFIDENCE VALUE >= 0.7(WITH NAME)

No	(C)	Consequent Name	Antecedent Name
1	0.94	Statistics	Statistics Practice
2	0.8	XML and Web Service Practice	XML and Web Service
3	0.8	Computer Network	Computer Network Practice
4	0.7	Object Oriented Programming	Database Design, Object Oriented Programming Practice
5	0.7	Discrete Mathematics	Calculus, Object Oriented Programming Practice
6	0.7	Database Design	Object Oriented Programming, Introduction to Information Technology
7	0.7	Database Management System	Calculus, Database Management System Practice
8	0.7	Calculus	Discrete Mathematics, Computer System
9	0.7	Calculus	Discrete Mathematics, Indonesian Language
10	0.7	Database Management System	Database Management System Practice
11	0.7	Discrete Mathematics	Object Oriented Programming, English II
12	0.7	Calculus	Indonesian Language, Database Management System
13	0.7	Calculus	Algorithm and Programming, Religion
14	0.7	Database Management System	Discrete Mathematics, Database Management System Practice
15	0.7	Discrete Mathematics	Database Design, Object Oriented Programming Practice
16	0.7	Object Oriented Programming	Discrete Mathematics, Object Oriented Programming Practice
17	0.7	Database Design	Object Oriented Programming, English II

From table 10, it is found that the association rule with the highest confidence value is MF511 –> MF142 (Statistics –> Statistics Practice) = 0.94%. This means that of all the students who retake courses Statistics, 94% of them took the statistics practice courses also.

This rule can help for the curriculum planners and academic directorate to manage or redesigning curriculum, changing teaching and assessment methodologies [7]. Example, the rule subject courses MF511 --> MF142 which has 93.5% confidence value, can be followed up by head of the study program to redesigning a class with same lecturer, or changing assessment methodologies. So that courses will get better percentage of student to pass.

IV. CONCLUSION

Some of point conclusion of this research:

- One of association rule implementation on University is to find relationship between the taking of one repeated course and other repeated courses.
- Support (s) value varies from 1% to 6%. While, confidence (c) value varies from 6% to 94%.
- If the support value threshold is smaller, there will be more association rule and.
- Statistics is course mostly repeated by students together with statistics practice with support value (s) 1% and confidence value (c) 94%.

REFERENCES

[1] Nugroho, Heru.”Conceptual Model Of It Governance For Higher Education Based On Cobit 5 Framework” (2013). Journal of Theoretical and Applied Information Technology.

[2] Lahtinen, Essi; Ala-Mutka, Kirsti; Jarvinen, Hannu-Matti, “A Study of the Difficulties of Novice Programmers” (2005). ITiCSE’05, Monte de Caparica, Portugal.

[3] Information System Directorat of Telkom University (Telkom Applied Science School Data)

[4] J.Han, Kamber M. “Data Mining Concepts and Techniques”. CA: Morgan Kaufmann, San Francisco.2011

[5] Assosiation Rule http://www.knime.org/files/nodedetails/_mining_subgroup_Association_Rule_Learner_Borgelt_.html

[6] Tan, Pang Nin; Steinbach, Michael; Kumar, Vipin. “Introduction To Data Mining” (2006). Addison-Wesley. ISBN 0-321-32136-7.

[7] Kumar Varun Dr; Chadha Anupama. “Mining Association Rules in Student’s Assessment Data”. IJCSI International Journal of Computer Science Issues, Vol 9, Issue 5, No 3, September 2012. ISSN (Online) : 1694-0814.

[8] Suriya S; Dr Shantarajah, S.P; Deeplakshmi. “A Complete Survey On Assosiation Rule Mining With Relevance to Different Domain”. International Journal of Advanced Scientific and Technical Research, Issue 2, Volume 1 (February 2012). ISSN : 2249-9954

[9] Shweta MS; Garg Kanwal Dr. “Mining Efficient Assosiation Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Assosiation Rule Algorithms”. International Journal Of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013. ISSN: 2277 128X

[10] Bhandari, Pranay; Rajeswari, K; Tonge, Swati; Shindalkar, Mahadev. “*Improved Apriori Algorihtms : A Survey*”. International Journal of Advanced Computational Engineering and Networking. ISSN: 2230-2106. Vol: 1- Issue-2, April. 2013