OPEN ACCESS
Indo-JC
Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 5, Issue. 2, September 2020. pp. 1-16
doi:10.21108/indojc.2020.5.2.434

Apriori Association Rule for Course Recommender system

Fakhri Fauzan #1, Dade Nurjanah #2, Rita Rismala #3

# School of Computing, Telkom University
Bandung, Indonesia

[1] fakhrifauzan@student.telkomuniversity.ac.id
[2] dadenurjanah@telkomuniversity.ac.id
[3] ritaris@telkomuniversity.ac.id

Abstract

Until recently, recommender systems have been applied in learning, such as to recommend appropriate courses. They are based on users' ratings, learning history, or curriculum that provide the relationship between courses. The last approach, however, can't be applied to Massive Open Online Courses (MOOCs) that don't maintain such information. Hence, course recommender systems for MOOCs must be based on other learners' experiences. This paper discusses such recommender systems. We apply the Apriori Association Rule and the case study used in this study is the Canvas Network dataset and the HarvardX-MITx dataset. The proposed recommender system consists of a pre-processing to normalize data and reduce anomalous data, data cleaning to handle empty data, K-Modes clustering to group users, grouping registration transactions for filtering user registration transactions, and finally, rule formation using the Apriori Association Rule. The performance of the association rules obtained, a lift ratio evaluation metric is used. The experiment results show the best parameters in this study are 0.01 for minimum support and 0.6 for minimum confidence. With these two parameters, the number of rules and the average lift ratio value on the Canvas Network dataset are 110 rules and 19.055, while the HarvardX-MITx dataset is 48 rules and 3.662.

Keywords: apriori association rules, courses, recommender system

Abstrak

Sampai saat ini, sistem rekomendasi telah diterapkan dalam pembelajaran, seperti merekomendasikan mata kuliah yang sesuai. Mereka didasarkan pada peringkat pengguna, histori pembelajaran, atau kurikulum yang menyediakan hubungan antar mata kuliah. Pendekatan terakhir, bagaimanapun, tidak dapat diterapkan pada *Massive Open Online Courses* (MOOCs) yang tidak menyediakan informasi tersebut. Oleh karena itu, sistem rekomendasi mata kuliah untuk MOOC harus didasarkan pada pengalaman peserta didik lainnya. Penelitian ini membahas sistem rekomendasi tersebut. Kami menerapkan *Apriori Association Rule* dan studi kasus yang digunakan pada penelitian ini adalah *Canvas Network dataset* dan *HarvardX-MITx dataset*. Sistem rekomendasi yang diusulkan terdiri dari *pre-processing* untuk menormalkan data dan mengurangi data yang tidak normal, data *cleaning* untuk menangani data kosong, pengklasteran K-Modes untuk mengelompokkan pengguna, pengelompokan transaksi registrasi digunakan untuk menyaring transaksi registrasi pengguna, dan terakhir pembentukan aturan menggunakan *apriori association rule*. Untuk menentukan performa aturan asosiasi yang diperoleh, digunakan metrik evaluasi *Lift Ratio*. Hasil percobaan menunjukkan bahwa parameter terbaik yang diperoleh dalam penelitian ini adalah 0,01 untuk minimum support, dan 0,6 untuk minimum confidence. Dengan kedua parameter ini, jumlah *rule* dan rata-rata nilai *lift ratio* pada *Canvas Network dataset* adalah 110 *rule* dan 19,055, sedangkan pada *HarvardX-MITx dataset* adalah 48 *rule* dan 3,662.

Kata Kunci: *apriori association rules*, mata kuliah, sistem rekomendasi

Received on Jan 2020. Accepted on Aug 2020

I. Introduction

Courses election is very influential during the study period. Students often need guidance in choosing courses as a condition for completing their studies [1]. Because taking courses is a student's authority, enough information about courses is required. In addition, understanding of self-capability must become another consideration in taking certain courses.

Until recently, course recommender systems have been applied in previous studies. Farzan and Brusilovsky [2] applied CourseAgent for course recommendations at the School of Information Sciences at the University of Pittsburgh, based on students' ratings, career goals, and feedback given by previous students [2]. The results of the study show that around 23% of students choose at least one recommended class. Other studies [3] built AACORN, a recommender system that applies case-based reasoning to graduate students at CTI DePaul [3]. AACORN recommends courses based on 4 features including student academic programs, curriculum terms, the overall average value of students, and the history of student registration. The study resulted in the percentage of relevant courses recommended reaching 80%.

Other studies used collaborative filtering approaches at The Indian Institutes of Management to predict marks that students will get in different courses based on their performance in previous courses [4]. The result is Mean Absolute Error (MAE) scores which are in the range of 0.33 to 0.38. Furthermore, another study built a tool, namely RARE, which applied association rules based on user preferences [1]. The advantage of RARE is that it can resolve cold start problems because the rule formation is done in an offline phase and the weights are always updated every time users give new feedback. The accuracy obtained by RARE reaches 90%. The study shows that the use of association rules makes a recommender system has a very intuitive framework in recommending items when there is an explicit or implicit transaction. Furthermore, it resulted in higher accuracy than a kNN collaborative filtering method [5]. The purpose of this study is to develop a system of course recommendations by using apriori association rules to assist students in determining the courses taken.

A. *Topics and Limitations*

In our research, we have developed a course recommender system of new courses to be taken by students in MOOCs. Some educational institutions provide a variety of MOOCs' services, such as Canvas Network, HarvardX, MITx, EdX, Udacity, Udemy and so on. The very rapid development of MOOCs has made a large number of course materials related to the offered courses and made it difficult for students to choose courses. Therefore, a tool that can recommend appropriate courses in MOOCs is needed.

The choice of the MOOC platform used in this study relates to the availability of published and anonymous dataset. Hence, the dataset used in this study is the Canvas Network Person-Course (1/2014 - 9/2015) De-Identified dataset and the HarvardX-MITx Person-Course dataset AY 2013. Canvas dataset consists of more than 325,000 records, and each record represents an activity of one user in one of 238 courses offered in the Canvas Network [6]. On the other hand, the HarvardX-MITx dataset consists of records of user activities in 13 courses offered in the edX platform in the first year [7].

The method used is Association Rules because previous studies showed its better performance than other methods, especially in overcoming cold start problems [1]. Research by Sunita and Lobo [8] made comparisons among four algorithms regarding association rules in the case of the course recommender systems. The four algorithms are apriori association rule, predictive apriori association rules, tertius association rules, and filtered associator. The results of these studies show that the apriori association rules algorithm resulted in the best performance since users agree with all the recommendation output. The problem covered in this study is how the implementation and performance of the course recommender system use the Apriori Association Rules, with case studies is the Canvas Network dataset and the HarvardX-MITx dataset.

The limitations of this study are due to the availability of the Canvas Network dataset and the HarvardX-MITx dataset. The Canvas Network dataset has a record of each activity in January 2014 to September 2015 period. On the other hand, the HarvardX-MITx dataset has records of user activities in the 2013 academic year (Fall 2012, Spring 2013, and Summer 2013). The normalized user data and the missing values were also ignored in this study. The purpose is to make the recommender system for courses more ideal.

Since it cannot be tested directly based on where this dataset was collected, the additional limitations of this research are carried out until the formation of the rules used for recommendations, and needed a metrics that can measure the correlation of the results of the recommendations obtained. So the metric used in this study is the lift ratio metric that can measure the correlation value of the rules obtained [9].

B. Purpose

The research aims to develop a course recommender system using Apriori Association Rule and analyze the performance of the recommender system, with the Canvas Network dataset and the HarvardX-MITx dataset as case studies.

II. Literature Review

A. Courses

According to the Indonesian Language Dictionary (KBBI), a course is a unit of learning taught at the college level. Courses are designed based on the curriculum with the aim that students can have knowledge and abilities by the majors/study programs taken.

B. Recommender System

The Recommender System is a software and technique to give suggestions about items that are considered useful to users [4]. These suggestions relate to various decision-making processes, such as what items to buy, what music you want to listen to or what online news to read. Some application of the recommender system, namely:

- Entertainment - recommendations for film, music, and IPTV.
- Content - newspaper personalization, document recommendations, web page recommendations, e-learning applications, and e-mail filters.
- E-commerce - product recommendations to buy such as books, cameras, computers.
- Services - travel service recommendations, expert recommendations for consultations, recommendations for rental homes.

To get recommendations that meet user preferences, recommender systems will actively collect various types of data about users and use the data in performing recommendations. In general, there are 3 main objects in the recommender system, namely item, which is an object that will be recommended, users, who get the recommendation results, and transaction, which records interactions between items and users. According to H. Drachsler et al. [10], the techniques commonly used in the recommender system are as follows.

1. *Content-based filtering*, users will be recommended with items similar to what they like in the past. The content-based recommender system will analyze a series of items and/or descriptions previously favored by users, and build models or profiles of user interests based on the features of the item.
2. *Collaborative Filtering*, users will be recommended with items that people like with similar tastes and preferences in the past. A collaborative filtering recommender system will predict user interest in new items based on recommendations from other people with similar interests.
3. *Demographic-based filtering* classifies users according to their profile attributes and makes recommendations based on demographic classes.
4. *Utility-based filtering*, make suggestions based on a calculation of utility of each item for a user, for whom the utility function must be stored.
5. *Knowledge-based filtering* shows items based on logical inferences about user preferences.
6. *Hybrid filtering*, combining two or more recommendation methods to get better performance and overcome the shortcomings of each method.

C. Association Rules

Association Rules is a method in data mining that focuses on searching for rules that can predict the appearance of an object in a transaction [5]. Examples of applications in everyday life when conducting a transaction at the convenience store are related to an item to other items that will be purchased at the same time. The role of association rules makes it easier to find the possibility of a buyer buying an item against another item. This is commonly called market basket analysis.

An itemset is defined as a set of one or more items. To make it easier to define the number of items in an itemset we can use the term k-itemset, where the k-value determines the number of items in an itemset. Examples of course itemset are {A, B, C} as 3-itemset, {D, E} as 2-itemset, and {F} as 1-itemset. The representation of association rules can be defined as X → Y, where X and Y are itemsets. For example, a student has taken courses D and E, then take the F course, it can be represented as follows:

{D, E} → {F} (*support* = 40%, *confidence* = 50%)

Giving a value of support of 40% and confidence of 50% to measure the interestingness of information, both of which show the interestingness and certainty of the rules built in the association rule. Support is defined as how much a rule applies to a data set or in other words, is the comparison of the occurrence of an item set to the overall item set, while for Confidence it is defined by how many items in Y appear in transactions containing X [11]. Values of support and confidence are in the range [0; 1] [12]. The formal definition of support and confidence is as follows.

$$Support, s(X \to Y) = P(X \cup Y) = \frac{\sigma(X \cup Y)}{N} = \frac{The\ Total\ Transaction\ contains\ X\ and\ Y}{Number\ of\ Transactions} \quad (1)$$

$$Confidence, c(X \to Y) = P(Y|X) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{The\ Total\ Transaction\ contains\ X\ and\ Y}{The\ Total\ Transaction\ contains\ X} \quad (2)$$

Based on these two formal definitions, the relationship between support and confidence with a course taking rule has meaning if $s(X \to Y)$ approaches one, then the appearance of $X \cup Y$ transactions is greater. Given a collection of T transactions, the purpose of association rule mining is to find all rules that have terms of *support* ≥ *minsup* and *confidence* ≥ *minconf* [11]. Minimum support or minsup is the minimum threshold for support. Therefore, if an item set has a support value below the specified threshold, then all the possible itemset will be pruned. On the other hand, minimum confidence or minconf is the minimum threshold for confidence. So, if there is a rule that has a confidence value below the specified threshold, then the rule is pruned. If a rule can meet these two conditions, then the rule can be said with a strong rule. To fulfill these two conditions, it can be done using the brute-force approach, however, this approach requires enormous computational costs. To overcome this, two steps are taken to describe the problem, namely:

1. ***Frequent Itemset Generation***, the purpose is to find all itemset that meets the minsup threshold. This item is called frequent itemset.
2. ***Rule Generation***, the aim is to extract all the rules that have confident values that are high from all frequent itemset obtained in the previous step. These rules are called strong rules.

D. Apriori Algorithm

Apriori algorithm is a method used to decrease the number of candidate items that searched during the frequent itemset generation process [11]. The principle of an apriori is that 'If the itemset included in the frequent itemset then the entire subset of an itemset also included in the frequent itemset'. To fulfill this principle, a support value used to prune the itemset candidate.

Figure 1 is an illustration of the apriori principle. If the itemset {c, d} included in the frequent itemset, then all subset of the itemset also included in the frequent itemset.

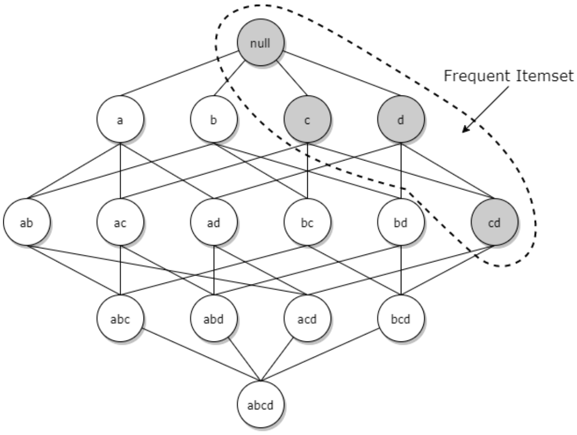

Figure 1. Illustration of the Apriori principle.

Figure 2 is an illustration of support-based pruning. If an itemset {c, d} is not included in the frequent itemset, then all supersets of {c, d} are also not included in the frequent itemset.

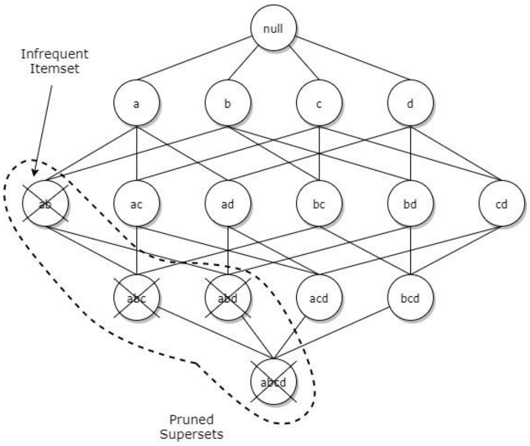

Figure 2. Support-based pruning illustration.

Figure 3 shows the apriori algorithm.

```
$L_1$ = {large 1-itemsets};
**for** (k=2; $L_{k-1} \neq 0$; k++) **do begin**
  $C_k$ = apriori-gen($L_{k-1}$); //Generate new candidates
  **forall** transactions $t \in \boldsymbol{D}$ **do begin**
    $C_t$ = subset($C_k$, $t$)
    **forall** candidates $c \in \boldsymbol{C_t}$ **do**
      c.count++;
    end
  $L_k = \{c \in C_k \mid c.\text{count} \geq \text{minsup}\}$
**end**
```

Figure 3. Apriori Algorithm.

III. RESEARCH METHOD

A. *System Overview*

Figure 4 is an overview flowchart of Course Recommender systems with the Apriori Association Rules.

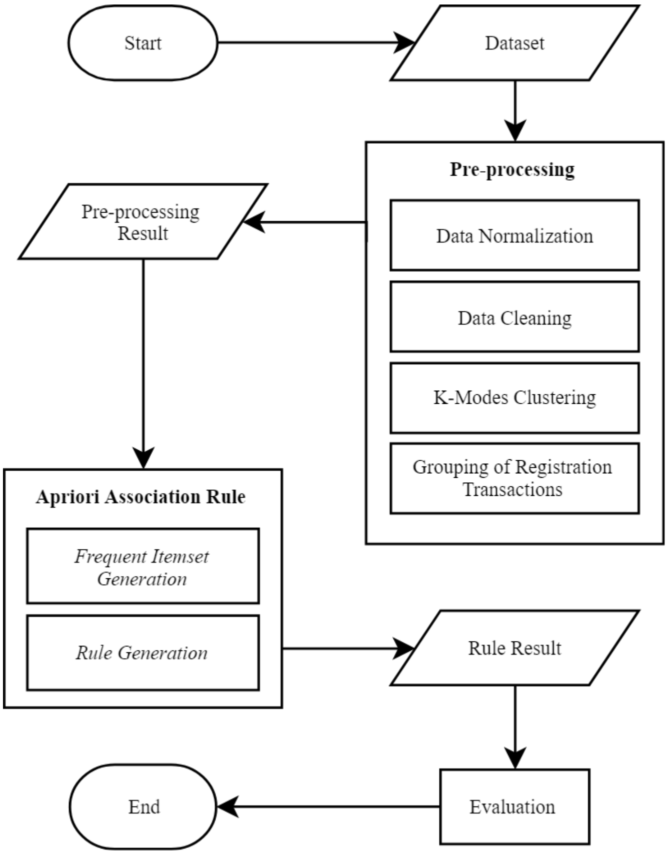

Figure 4. System Overview.

B. *Dataset*

In this study there are 2 datasets used, namely:

1) Canvas Network dataset

Canvas Network dataset or Canvas Network Person-Course (1/2014 - 9/2015) De-Identified Open dataset is a dataset published by the Canvas Network (Instructure) on the page https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/1XORAL under the CC-BY 4.0 license. This dataset consists of data on the Canvas Network course that runs from January 2014 - September 2015. This dataset includes more than 325,000 records, and each record represents the activities of one person in one of the 238 courses available. The data structure used in this dataset based on the HarvardX-MITx Person-Course 2014. Table 1 shows several courses available based on disciplines in the period of January 2014 - September 2015.

Table 1. Course count by discipline in this Canvas Network dataset [6].

Discipline	Number of Courses
Professions and Applied Sciences	73
Education	57
Humanities	29
Business and Management	28
Interdisciplinary & Other	16
Computer Science	9
Social Sciences	8
Mathematics & Statistics	7
Physical Sciences	6
Total	**238**

Each table in the data set records registration data for one course, while each row in the data set records a registration of a student in a course. So, if one person registered in 3 courses during the period covered by the dataset, then that person has three rows associated with the personal user ID. Table 2 is the attribute found on the Canvas Network dataset.

Table 2. Canvas Network dataset attribute.

Attribute	Description	Attribute	Description
course_id_DI	a unique identifier for the course	age_DI	age brackets
userid_DI	a unique identifier for the user	gender	-
Registered	the status of registered for the course	start_time_DI	quarter and year that the first user interaction occurred
Viewed	the number of interactions within the course is greater than 1	course_start	quarter and year that the course officially started
Explored	the user interacted with or viewed >=50% of the course modules	course_end	quarter and year that the course officially ended
completed_%	percent of total required content modules completed	last_event_DI	quarter and year that the last user interaction occurred
course_reqs	content module status (if >= 3) with requirements	nevents	count of distinct interactions with the course
grade_reqs	assignments status in course (if >= 3)	ndays_act	count of distinct days with one or more events
primary_reason	standardized reason for taking a course	nforum_posts	number of posts total in discussion forums throughout the course
final_cc_cname_DI	-	course_length	number of days that course officially ran or that course had participant activity
learner_type	standardized type of learner	grade	the final grade in the course as a percent
expected_hours_week	standardized range of hours per week	discipline	a generalization of the course title as the discipline of the course
LoE_DI	the highest level of education completed	ncontent	-

2) HarvardX-MITx dataset

HarvardX-MITx dataset or HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset is a dataset published by MITx and HarvardX on the https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147. This dataset has data from the first year (2013 Academic Year: Fall 2012, Spring 2013, and Summer 2013) from MITx and HarvardX courses on the edX platform. Each record represents the activities of one person in one edX course. There are 13 courses in the dataset, consisting of 8 courses from MITx and 5 courses from HarvardX, Table 3 lists all courses available in the HarvardX-MITx dataset.

Table 3. HarvardX-MITx dataset course [7].

Institution	Course Code	Short Title	Full Title
HarvardX	CB22x	HeroesX	The Ancient Greek Hero
HarvardX	CS50x	-	Introduction to Computer Science I
HarvardX	ER22x	JusticeX	Justice
HarvardX	PH207x	HealthStat	Health in Numbers: Quantitative Methods in Clinical & Public Health Research
HarvardX	PH278x	HealthEnv	Human Health and Global Environmental Change
MITx	14.73x	Poverty	The Challenges of Global Poverty
MITx	2.01x	Structures	Elements of Structures
MITx	3.091x	SSChem	Introduction to Solid State Chemistry
MITx	6.002x	Circuits	Circuits and Electronics
MITx	6.00x	CS	Introduction to Computer Science and Programming
MITx	7.00x	Biology	Introduction to Biology the Secret of Life
MITx	8.02x	E&M	Electricity and Magnetism
MITx	8.MReV	MechRev	Mechanics Review

This dataset is at the per-person one-row level, per course defined in 1 table. So, if one person registered in 3 courses during the period covered by the dataset, then that person has three rows associated with the personal user ID. Table 4 is an attribute of the HarvardX-MITx dataset.

Table 4. HarvardX-MITx dataset attribute.

Attribute	Description	Attribute	Description
course_id	identifies institution, course name, and semester	gender	-
userid_DI	a unique identifier for the user	grade	the final grade in the course
registered	the status of registered for the course	start_time_DI	date of course registration
viewed	access status for the 'Courseware' tab (video, problem sets, exams)	last_event_DI	date of last interaction with course
explored	access status at least half of the chapters in the courseware	nevents	number of interactions with the course
certified	anyone who earned a certificate	ndays_act	number of unique days student interacted with course
final_cc_cname_DI	-	nplay_video	number of play video events within the course
LoE_DI	the highest level of education completed	nchapters	number of chapters with which the student interacted
YoB	year of birth	nforum_posts	number of posts to the Discussion Forum
roles	identifies staff and instructors	incomplete_flag	identifies records incomplete course

C. Pre-processing

Pre-processing is made to process raw data into data used in the system being built. Several stages carried out on pre-processing, namely data normalization, data cleaning, and K-modes clustering.

1) Data Normalization

Both datasets are available in one table with abnormal form since there is a recurring key or duplication of the primary key. A normalization process is needed to drop duplicate data and make the relationship between entities/tables clear. The normal form targeted is the second normal form (2NF) where each non-key attribute functionally depends on the primary key. Normalization is carried out based on the conditions contained in the second normal form, by dividing the two datasets into three main entities/tables, namely User, Course and Registration. Example of normalizing a user in the canvas network dataset shown in Table 5.

Table 5 Example of Data Normalization

Canvas Network Dataset	
Before	After
course_id_DI: 832945550 discipline: 'Social Sciences' userid_DI: 832300004 registered: 1 viewed: 1 explored: 0 grade: 0.8220000000000001 grade_reqs: 1 completed_%: nan course_reqs: 1 final_cc_cname_DI: '*' primary_reason: 'I hope to gain skills for a new career' learner_type: 'Active participant' expected_hours_week: 'Between 1 and 2 hours' LoE_DI: "Master's Degree (or equivalent)" age_DI: '{55 or older}' gender: '{}' start_time_DI: '2015 Q1' last_event_DI: '2015 Q2' nevents: 355.0 ndays_act: 10.0 ncontent: 14.0 nforum_posts: 16.0 course_length: 27	**Table User** userid_DI: 832300004 age_DI: '{55 or older}' LoE_DI: "Master's Degree (or equivalent)"
	Table Course course_id_DI: 832945550 discipline: 'Social Sciences' course_start: '2015 Q1' course_end: '2015 Q2' course_length: 27
	Table Registration course_id_DI: 832945550 userid_DI: 832300004 registered: 1 viewed: 1 explored: 0 grade: 0.8220000000000001 grade_reqs: 1 completed_%: nan course_reqs: 1 primary_reason: 'I hope to gain skills for a new career' learner_type: 'Active participant' expected_hours_week: 'Between 1 and 2 hours' start_time_DI: '2015 Q1' last_event_DI: '2015 Q2' nevents: 355.0 ndays_act: 10.0 ncontent: 14.0 nforum_posts: 16.0

Table 6 is the result of data normalization in both datasets.

Table 6. Data Normalization Result.

Table	dataset	
	Canvas Network	HarvardX-MITx
User	userid_DI, age_DI, LoE_DI	userid_DI, gender, YoB, age_DI, LoE_DI, final_cc_cname_DI
Course	course_id_DI, discipline, course_start, course_end, course_length	course_id, course_start, course_end
Registration	course_id_DI, userid_DI, registered, viewed, explored, grade, grade_reqs, completed_%, course_reqs, primary_reason, learner_type, expected_hours_week, start_time_DI, last_event_DI, nevents, ndays_act, ncontent, nforum_posts	userid_DI, course_id, certified, grade, start_time_DI, last_event_DI, nevents, ndays_act, nplay_video, nchapters, nforum, posts, incomplete, flag

2) Data Cleaning

At this stage, handling missing values, inconsistent and irrelevant data on all tables that formed during the normalization stage is carried out. Data cleaning will affect the complexity and performance of the built tool because the data used is smaller and more relevant. At this stage, the table that greatly affects the amount of data is from the user table. Because the user table is the main determinant of the amount of data in the registration table, if there is user data filtered from the user table, then the user data is automatically not available on the registration data. Table 7 shows that the number of records in HarvardX-MITx is higher than in the Canvas Network dataset. Furthermore, the percentage of unique data datasets and filtered data in Canvas Network has a small size. The filtered data are unique and they resulted from a data cleaning activity. The facts about the data indicate that the Canvas Network dataset has more missing values than the HarvardX-MITx dataset. In terms of courses offered, there are 13 courses recorded in the HarvardX-MITx dataset; it is fewer than the Canvas Network dataset which has 238 courses. As a result of data pre-processing, data records of registration used in rule formation consist of 69502 records in Canvas Network dataset registration and 500829 records in the HarvardX-MITx dataset.

Table 7. Data Cleaning Result.

dataset	Table	Amount of data		Percentage (%)
		Unique	Filter	
Canvas Network dataset	User	224914	31201	13.87 %
	Course	238	238	100 %
	Registration	325199	69502	21.37 %
HarvardX-MITx dataset	User	476532	383335	80.44 %
	Course	13	13	100 %
	Registration	641138	500829	78.11 %

3) K-Modes Clustering

The K-Means clustering algorithm can be applied to numeric type attributes, by calculating the average score of parameters as a distance between objects, thus it is not possible to apply the algorithm to nominal/categorical data types [13][14]. To cluster data based on nominal/categorical data, a modified K-Means clustering algorithm called the K-Modes algorithm, is needed. Unlike the K-Means algorithm that uses average scores, the K-Modes algorithm uses mode scores, taken from the values of the parameter which mostly appeared. Changes made by the K-Modes algorithm to the K-Means algorithm are dissimilarity, mode, and frequency-based methods [13].

Since some attributes in the datasets we used are nominal, we apply a K-Modes clustering algorithm. Before conducting experiments, the HarvardX-MITx dataset needs a categorization process to convert 'YoB' attribute to an age categorization as found in the Canvas Network dataset as 'age_DI', so that at the clustering stage K-Modes balanced in both datasets because the attributes used are the same. The attributes used in the clustering process are 'age_DI' and 'LoE_DI'. To find the number of clusters or the best *k* value using the Elbow method. The Elbow method has a basic idea which is increasing the number of clusters to decrease in-cluster variance of all existing clusters [15] or choosing the number of clusters when there are significant changes in the value of distortion or sum of squared error (SSE), which is followed by changes in the value of the distortion which tends to be stable.

After determining the value of *k* with the Elbow method, the next is grouping user data into two datasets. The clusters are stored in a column labeled 'labels'. Table 8 and Table 9 show some examples of the results of the user data clustering process using K-Modes.

Table 8. Canvas Network dataset User Clustering Results.

userid_DI	age_DI	LoE_DI	labels
832300004	{55 or older}	Master's Degree (or equivalent)	4
832300077	{19-34}	Completed 4-year college degree	3
832300082	{19-34}	Master's Degree (or equivalent)	0
832300105	{34-54}	Some college, but have not finished a degree	8
832300112	{34-54}	Ph.D., J.D., or M.D. (or equivalent)	10

Table 9. HarvardX-MITx dataset User Clustering Results.

userid_DI	gender	YoB	age_DI	LoE_DI	final_cc_name_DI	labels
MHxPC130000002	f	1990	{19-34}	Secondary	Canada	8
MHxPC130000003	m	1991	{19-34}	Secondary	Brazil	8
MHxPC130000004	m	1993	{19-34}	Secondary	India	8
MHxPC130000006	m	1975	{34-54}	Bachelor's	United States	5
MHxPC130000007	f	1968	{34-54}	Master's	United States	1

4) Grouping Registration Transactions

After completing user clustering using K-Modes, the transaction history for each user cluster will be grouped. Since a rule must be made from two or more items [16], transaction data are filtered to get data of users who have done two registration. The format performed from this process is a list of lists, each record in the list of user links to a list of *course_id* contained with the id of course that has been taken by the user.

D. *Apriori Association Rules*

The formation is done using the Apriori Association Rules method. There are 2 main stages in the process of forming a rule, namely frequent item generation, and rule generation. Output at this phase is the Result of the Formation of Rules. A rule defined as *if x then y*, where *x* is the antecedent and *y* is the consequent. In the rule generation stage, the calculation of the values of support, confidence, and lift ratio is carried out on each rule. The minimum support and minimum confidence values are determined in a simulation; the formed rule must have support and confidence scores higher than the minimum scores. The stage of rule formation is carried out on each cluster of users, thus each cluster will have different rules. Table 10 shows some examples of frequent item generation calculations for minimum support = 0.2 and minimum confidence = 0.6. Furthermore, Table 11 shows the results of rule formation applied to cluster 5 of the HarvardX-MITx dataset. The orange highlighted part are pruned during the process because their support scores are less than the minimum support.

Table 10. Frequent Item Generation Process

1-itemset	support
{6.00x}	0.533
{CS50x}	0.604
{ER22x}	0.211
{PH278x}	0.186
{8.MReV}	0.018

2-itemset	support
{6.00x, CS50x}	0.426
{6.00x, ER22x}	0.032
{CS50x, ER22x}	0.059

Table 11. Rule Generation Results.

Antecedents	Consequent	Support	Confidence	Lift
{6.00x}	{CS50x}	0.426083	0.798412	1.320224
{CS50x}	{6.00x}	0.426083	0.704554	1.320224

E. *Lift Ratio Evaluation*

To evaluate the rules formed, a lift ratio value is used to measure the interestingness of a rule. Lift Ratio shows the level of strength in an association rule or can be called a measure of simple correlation on a rule. The Lift Ratio value is in the range [0; +∞), thus the higher the value of lift ratio indicates a stronger association rule [17]. The lift ratio is defined in the following equation [18].

$$lift(X \rightarrow Y) = \frac{P(X \cup Y)}{P(X)P(Y)} = \frac{sup(X \cup Y)}{sup(X)\ sup(Y)} = \frac{conf(X \rightarrow Y)}{sup(Y)} \tag{3}$$

Based on the lift ratio values found in the above equation, it can be defined as follows [9]:

– If the lift ratio is less than 1, then the appearance of X is negatively correlated with the appearance of Y, which means that the appearance of one might lead to the absence of the other.
– If the lift ratio more than 1, then X and Y are positively correlated, meaning that the occurrence of one indicates the other’s occurrence.
– If the lift ratio equals 1, then X and Y are independent and there is no correlation between them.

IV. Results And Discussion

A. *K-Modes Clustering Results*

To determine the number of clusters used in each dataset, some simulations were conducted to compare the value of distortion or the sum of squared error (SSE) for each value of *K*. Then the value of *K* used is determined based on the Elbow method. Clustering simulations are carried out using *K* values *starting* *K* = 2 to *K* = 20 in both datasets. Figure 5 is a graph of the comparison of the value of distortion or the SSE for the two datasets.

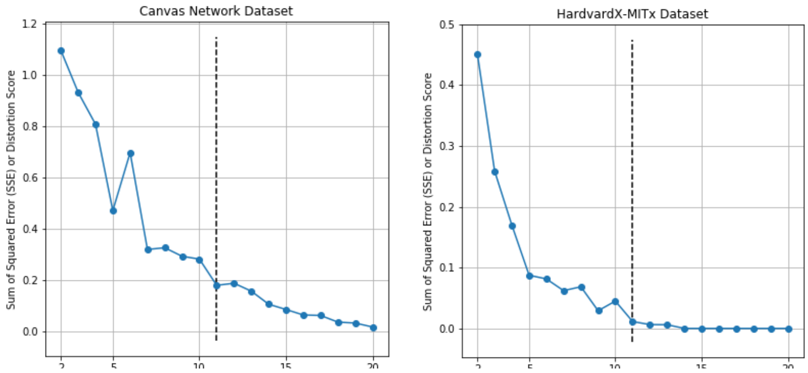

Figure 5. Graph Comparison of SSE Values for each dataset User Cluster.

Based on the graph in Figure 5, significant changes in the distortion values and followed by changes in the distortion value that tend to be stable are found in *K* = 11 for the Canvas Network dataset, and *K* = 11 for the HarvardX-MITx dataset. The value of *K* is used to determine the number of cluster users. Table 12 and Table 13 are the results of the user clustering process in each dataset.

Table 12. Canvas Network dataset Preprocessing Results (K = 11).

Canvas Network dataset			
Cluster-	Number of		Ratio
	User	Transaction	(%)
1	6661	2547	38.23
2	1439	581	40.37
3	5666	2362	41.68
4	4294	1685	39.24
5	3515	1332	37.89
6	1897	749	39.48
7	1227	503	40.99
8	846	372	43.97
9	1342	553	41.20
10	2998	1259	41.99
11	1316	517	39.28

Table 12 shows the highest number of users on the Canvas Network dataset is in cluster 1, which contains 6661 users. This is also directly proportional to the highest number of transactions in cluster 1, which contains 2547 registration transactions. However, the highest ratio is reached in cluster 8 with a ratio of 43.97%, which means that 43.97% of users have taken two courses or more.

Table 13. HarvardX-MITx dataset Preprocessing Results (K = 11).

HarvardX-MITx dataset			
Cluster-	Number of		Ratio
	User	Transaction	(%)
1	11206	1694	15.11
2	29293	4785	16.33
3	140123	32945	23.51
4	5281	766	14.50
5	1103	71	6.43
6	25033	4248	16.96
7	4107	431	10.49
8	5544	796	14.35
9	100469	25153	25.03
10	58486	12747	21.79
11	2690	280	10.40

Table 13 shows that the highest number of users on the HarvardX-MITx dataset is reached in cluster 3, which consists of 140123 users. It is conformance to the highest number of transactions in cluster 3, namely 32945 registration transactions. However, the highest ratio is reached in cluster 9 with a ratio of 25.03%.

Based on the two analyses above, some differences occurred that the two datasets have different user characteristics. It can be observed from the number of users and the ratio of clustering results. As we have explained previously that the number of users in the Canvas Network dataset is 31201 users and in the HarvardX-MITx dataset is 383335 users, the ratio of the number of users of both datasets is 8.13%. It is inversely proportional to the value of the ratio obtained. As shown in Table 12, the Canvas Network dataset gave ratio values which tend to be more evenly distributed in the range of 37.89 - 43.97%, while the ratio given by using the HarvardX-MITx dataset is in the range of 6.43 - 25.03%.

B. Minimum Support and Confidence Test Results for Rule

After determining the number of clusters in each dataset. The next stage is the rule generation using the Apriori Association Rules for each cluster generated. To get a comprehensive result of the rules formed, some simulations using different minimum support and minimum confidence values are applied to each rule generation in the user cluster. The minimum support and confidence values used in this simulation are 0.01, 0.02, 0.04, 0.06, 0.08, 0.10 and 0.2, 0.4, 0.5, 0.6, 0.8, 1.0. The selection of minimum support values is based on the availability of data resulting from the grouping of registration, while the minimum confidence value is determined based on the desired level of trust at the time of making the rules. Table 14 and Table 15 shows the results of experiments using some minimum values of support and minimum confidence in the number of rules formed. They show that the minimum support and minimum confidence scores greatly influences the number of rules formed.

Table 14. The number of Rule Canvas Network dataset.

Canvas Network dataset						
minsup/minconf	**0.2**	**0.4**	**0.5**	**0.6**	**0.8**	**1.0**
0.01	728	274	195	110	37	5
0.02	89	52	42	30	13	0
0.04	22	21	19	13	10	0
0.06	14	14	12	9	7	0
0.08	10	10	8	7	5	0
0.10	6	6	5	4	3	0

Table 14 shows the number of rules formed in the simulation on the Canvas Network dataset. The maximum number of rules, 728, is reached when the minimum value of support is 0.01 and the minimum confidence is 0.2. When the minimum confidence is 1.0, no rules are formed.

Table 15. The number of Rule HarvardX-MITx dataset.

HarvardX-MITx dataset						
minsup/minconf	**0.2**	**0.4**	**0.5**	**0.6**	**0.8**	**1.0**
0.01	462	142	94	48	25	23
0.02	321	100	54	24	2	0
0.04	192	77	43	20	2	0
0.06	123	64	42	20	2	0
0.08	71	45	34	20	2	0
0.10	43	36	29	19	2	0

Furthermore, Table 15 shows the number of rules formed at the HarvardX-MITx dataset. The maximum number of rules, 462 rules is reached when the minimum value of support is 0.01 and the minimum confidence value is 0.2. When the minimum confidence is 1.0, no rules are formed.

Based on the two results above, the two datasets have similarities in terms of the number of rules formed. Both datasets have the highest number of rules when the minimum value of support is 0.01 and the minimum confidence value is 0.2. In addition, when the minimum values support is 0.02, 0.04, 0.06, 0.08, 0.10 and minimum confidence 1.0, both datasets do not result in any rule. Hence, the parameter is not used as a parameter of the proposed course recommender system. Another finding is shown in Table 14 and 15 is that the greater the minimum values of support and confidence, the fewer the number of rules are formed. Hence, it can be concluded that the number of rules formed is strongly influenced by the minimum scores of support and confidence.

C. Minimum Support and Confidence Test Results for Lift Ratio

To measure the performance of the rules that have been obtained in the previous section, an evaluation was carried out in the form of calculating the lift ratio average in all user clusters to determine the correlation of the association rules formed. Evaluation is carried out by conducting a comprehensive simulation of all the rules formed by using different minimum values of support and confidence. Table 16 and Table 17 are the results of experiments using several minimum values of support and confidence with the average lift ratio.

Table 16. Average Lift Ratio for Canvas Network dataset.

Canvas Network dataset						
minsup/minconf	**0.2**	**0.4**	**0.5**	**0.6**	**0.8**	**1.0**
0.01	9.803	15.106	17.429	19.055	13.578	7.525
0.02	8.032	9.570	10.268	11.320	7.392	0
0.04	5.797	5.806	5.806	5.797	5.823	0
0.06	3.097	3.097	3.097	3.097	3.097	0
0.08	1.732	1.732	1.732	1.732	1.732	0
0.10	1.175	1.175	1.175	1.175	1.175	0

Table 16 shows the average lift ratio of the association rules formed in simulations using the Canvas Network dataset. From all the simulations, there were 31 simulations that produced an average lift ratio greater than one and the formed association rules have positive correlations. In addition, five simulations produced lift ratios equal to 0 because there are no association rules formed, as shown in Table 14. The highest average lift ratio, which is 19.055, is reached when the minimum value of support is 0.01 and the minimum confidence is 0.6.

Table 17. Average Lift Ratio for HarvardX-MITx dataset.

HarvardX-MITx dataset						
minsup/minconf	**0.2**	**0.4**	**0.5**	**0.6**	**0.8**	**1.0**
0.01	2.571	3.216	3.289	3.662	2.391	2.371
0.02	1.393	1.680	1.742	1.678	0.246	0
0.04	1.396	1.601	1.613	1.525	0.246	0
0.06	1.493	1.621	1.620	1.525	0.246	0
0.08	1.456	1.583	1.561	1.525	0.246	0
0.10	1.445	1.487	1.494	1.508	0.246	0

Table 17 shows the average lift ratio of the association rules formed in simulations using the HarvardX-MITx dataset. From all the simulations, there were 26 simulations that produced an average lift ratio greater than 1.5 and the formed association rules have positive correlations. In addition, five simulations produced association rules with a negative correlation since the lift ratio is between 0 and 1. Furthermore, the other five simulations produced lift ratios equal to 0 because there are no association rules formed, as shown in Table 15. The highest average lift ratio, which is 3.66, is reached when the minimum value of support is 0.01 and the minimum confidence is 0.6.

In contrast, the two datasets show differences in the distribution of the average lift ratio. Table 17 shows the average lift ratio value of the HarvardX-MITx dataset tends to be more evenly distributed in the range of 0-3.662, in comparison with the average lift ratio value of the Canvas Network dataset in the range 0-19.055. Some simulations produce an average lift ratio which is more than one, which means that the association rules formed have a positive correlation.

The highest average lift ratio value in the two datasets means that both conditions have a very high positive correlation value because it has a lift ratio value of more than 1 (based on equation 3) of the established rules, so the minimum support and minimum confidence values are contained in both simulations can be used in a recommendation system that can produce rule recommendations that can have the highest correlation value for each dataset.

The *minsup* and *minconf* parameters used in such simulations can be used as references in the proposed recommender system. However, based on the above results, there are variations in the results of the average lift ratio values greater than one that the Canvas Network dataset resulted in the list ration in the range of 1.175-19.055, while HarvardX-MITx dataset in the range of 1.393-3.662. These differences indicate that the greater the value of the lift ratio, the more ideal the recommended results of the courses obtained.

V. Conclusion And Future Work

Based on the results of testing and evaluation, it was concluded that the Apriori Association Rules method is well used in the course recommender system. The research finds is that the best parameter obtained is the minimum value of support, which is 0.01, and the minimum value of confidence, which is 0.6. With these two parameters, the Canvas Network dataset produces 110 rules and the average lift ratio is 19.055, while the HarvardX-MITx dataset produces 48 rules and the average lift ratio is 3.662. The research finds that the difference in minimum support and minimum confidence does not necessarily mean that the lift ratio values are smaller if the minimum support and minimum confidence values are greater.

A suggestion for further research on the course recommender system is to use a variety of Association Rule algorithms such as Apriori Hybrid, FP-growth and so on. The data used can be varied, the use of local university data will be advantageous for local universities. As a system improvement, we can evaluate the students' grades obtained for the taken courses. Furthermore, to get the results of better association rules, it is expected to use a more varied association rules measurement metric such as conviction, leverage, and coverage and so on.

References

[1] N. Bendakir and E. Aimeur, "Using association rules for course recommendation," *Proc. AAAI Work. Educ. Data Min.*, vol. WS-06-05, pp. 31–40, 2006.

[2] R. Farzan and P. Brusilovsky, "Encouraging user participation in a course recommender system: An impact on user behavior," *Comput. Human Behav.*, vol. 27, no. 1, pp. 276–284, 2011.

[3] R. Burke, "Aacorn: A CBR recommender for academic advising," 2005.

[4] S. Ray and A. Sharma, "A collaborative filtering based approach for recommending elective courses," *Commun. Comput. Inf. Sci.*, vol. 141 CCIS, pp. 330–339, 2011.

[5] F. Ricci, L. Rokach, and B. Shapira, *Recommender Systems Handbook*. 2011.

[6] C. Network, "Canvas Network Person-Course (1/2014 - 9/2015) De-Identified Open Dataset." Harvard Dataverse, 2016.

[7] Mit. and HarvardX, "HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0." Harvard Dataverse, 2014.

[8] S. B. Aher and L. M. R. J. Lobo, "A Comparative Study of Association Rule Algorithms for Course Recommender System in E-learning," 2012.

[9] J. P. Jiawei Han, Micheline Kamber, *Data Mining – Concepts & Techniques*. 2011.

[10] N. Manouselis, H. Drachsler, K. Verbert, and E. Duval, *Recommender Systems for Learning*. Springer-Verlag New York, 2013.

[11] D. P.-N. Tan, D. M. Steinbach, D. A. Karpatne, and D. V. Kumar, *Introduction to Data Mining (Second Edition)*, 2nd Editio. New York, NY: Pearson Education, 2018.

[12] L. M. Sheikh, B. Tanveer, and M. A. Hamdani, "Interesting measures for mining association rules," in *8th International Multitopic Conference, 2004. Proceedings of INMIC 2004.*, 2004, pp. 641–644.

[13] Suyanto, *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung, Indonesia: Informatika Bandung, 2017.

[14] S. B. Aher and L. M. R. J. Lobo, "Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data," *Knowledge-Based Syst.*, vol. 51, pp. 1–14, 2013.

[15] Suyanto, *Machine Learning: Tingkat Dasar dan Lanjut*. Bandung, Indonesia: Informatika Bandung, 2018.

[16] M. Awad and R. Khanna, *Machine Learning in Action*. 2015.

[17] M. Hahsler, B. Grün, and K. Hornik, "Introduction to arules - Mining Association Rules and Frequent Item Sets," *October*, pp. 1–28, 2006.

[18] W.-Y. Lin, M.-C. Tseng, and J.-H. Su, "A Confidence-Lift Support Specification for Interesting Associations Mining," pp. 148–158, 2007.