International Journal of Soft Computing 12 (5-6): 287-293, 2017
ISSN: 1816-9503

Graduate Rate Analysis of Student Using Data Mining and Algorithm Apriori

Setiawan Awan and Rusmawan Dadan
Department of Informatics Engineering, Langlangbuana University, Bandung, Indonesia

Abstract: The rapid growth of the accumulation of data has created a data-rich condition but minim of information. Data mining is mining process extracting information by getting the pattern or specific rules from large amounts of data that can solve that condition. Taking advantage of student data and student graduate data expected to yield information about the relationship with a graduate rate of the student holding the data through data mining technique. The graduate rate is measured from the time of study and GPA. Data mining techniques use apriori algorithm the information displayed in the form of support and confidence values from each graduate rate category.

Key words: Data mining, Apriori algorithm, graduate rates, student master data, information, condition

INTRODUCTION

Very required on the human's daily life, so that information will become one of the important element in the growth of the society now a days and on the future. However, the high information needs sometimes not balance with the presentment of the satisfy information, frequently that information still need to be excavated again from large amounts of data. The ability of information technology to collecting and saving various type of data far left its ability to analyze to summarize and to extract knowledge from data. The traditional methods for analyzing existing data, can't handle a large amount of data.

Utilization of existing data in information system to support decision-making activities not enough to just rely on operational data only, required an analysis of data for exploring the potential of existing data. The decision makers are trying to abuse of the data warehouse that already owned to excavated the useful information for helping to make a decision this is led to the emergence of a new subdivision of science for solving a problem of extracting information or important or interesting pattern from the large amounts of data the so-called data mining. The use of data mining techniques is expected can provide a knowledge previously hidden in the data warehouse to be a valuable information.

A college now a days required to have an eminence contend by utilizing all available resources. In addition to the resources of facilities, infrastructure and human an information system is one of the resources can be used for increase eminence contend. The information system can be used for processing and spread information to support the operational activities routine at once to support strategic decision making.

About study time of undergraduate (SI) regular mentioned that the period of normal study for bachelor's degree education consist by Anonymous (2013):

- The first stage of the year is scheduled in 2 semesters or 1 year
- The undergraduate stage is scheduled in 6 semesters or 3 years, after the altogether arrangement stage

So, the total of the period of study is 8 semesters, whereas extra-time study who permitted/qualify, would not result in the entire study period exceeds the maximum as follows:

- The 4 semesters or 2 years for the first stage of the year
- The 12 semesters or 6 years for the first stage of the year and undergraduate stage

With the study load at least 144 credit hours (semester credit unit credit semester) and at the most 160 credit hours. Based on April 2013 graduate data, 360-639 participants of bachelor's degree graduate at Institut Teknologi Bandung take >8 semesters study period. This shows that there still many students of bachelor's degree program at ITB take a period of study more than 8 semesters of the scheduled 8 semesters. Therefore, with utilizing student master data and student graduate data, the information of student graduate rate can be seen through data mining technique.

Contribution and research output: Student graduate rate can be seen from the period of study and IP

Corresponding Author: Setiawan Awan, Department of Informatics Engineering, Langlangbuana University, Bandung, Indonesia

(Indeks Prestasi), Grade Point Average (GPA) contained in the student graduate data. Data mining be expected can help to give information about student graduate rate with use student graduate data and student master data.

The problem is a lack of application for the decision-maker at ITB especially at education directorate to see information about the connection of graduate rate with student master data with data mining technique. The information in the form of support value and confidence the relation between graduate rate with the student master data.

Purpose of research: The goals from this research are generating an application to get a useful information about student graduate rate with data mining technique.

As for some benefits which are expected from this research is with this application be expected can helping to give information about the relationship between graduate rate with the student master data. Part of ITB can know theirs student graduate rate and find out the graduate rate affecting factors.

MATERIALS AND METHODS

The methods in data collection of this research is:

- Direct observation methods; direct observation at education directorate ITB to get the needed data
- Interview methods; conduct interviews with the parties that are directly related with the issues which are being discussed in this study to get a description and fundamental explanation
- Study literature methods; a source that can be used as a reference from the data source or the literature
- Browsing methods; collecting the reference which is sourced from the internet

RESULTS AND DISCUSSION

Research outcomes

Business understanding: In this research will searching for support value and confidence from the relationship between graduate rate with student master data. Not at all student master data will find the relation with graduation data, just a several attribute which approximately useful and its spread not too random. Because are too random data will make mining process takes a long time and even low-rate connection. Student master data to be searched the relationship is entry process, origin school, origin school city and study program. As for who will be processed.

The relationship between graduate rate with entry process. The result from this mining process can helping to find out how far USM and SNMPTN successful rate. The relationship between graduate rate with origin school and entry process

From entry process attribute and origin school to be searched the relation between graduate rate with origin school that the entering process through USM with the expectation can determine the rate of student successful with the certain school.

The relationship between graduate rate with birth city The relationship between graduate rate with origin city useful to know which region who have a high rate of successfull or low. With assumed that birth place is student's hometown.

The relationship between graduate rate with study program from a study, program attribute can find out the relationship between graduate rate and study program to know the graduate rate of the study program.

Data understanding: In this research, the data that used consist of two sources of data that is student master data and graduate data.

Student master data: Student masters data is student data recorded when the first time student entry into the university after re-registration. The data that recorded is personal identity and the student's origin school identity. Collection process performed at institute rate.

Graduate data: Graduate data is student's data who have passed. The recorded data is student's identity and the graduation completeness.

The student master data that taken in the sample are student data from 2006-2009 generation. This is based the needed of data who will connected with graduate data, with assumed that the student on 2006-2009 generation will graduate from 2010-2013 ranges time. However, graduate data that taken in is graduate data from 2010-2013. Both of that data obtainable from data based system information ITB academical. The taken data just from bachelor's degree student (S1).

Data preparation: In this study has to search for some attribute relationship from student master data with the graduate rate. Because not at all tables are used then needs a data purge in order to get a required relevant data. This purge is really important to increase performance in the mining process. The purge way is with deleting the unused attribute and deleting incomplete data, the used attribute consists of an attribute in graduate data and in student master data. The used attribute in student master data is:

- NIM attribute be used as a primary key to connecting with graduate data
- Entry process attribute be used for mining process in order to find out a relationship between graduate rate with entry lane that student use
- Origin name of school attribute is used for mining process in order to find out a relationship between graduate rate with origin school
- Birth place attribute be used for mining process in order to find out a relationship between graduate rate with student hometown

The used attribute in graduate data is:

- NIM be used as a primary key to connecting with student master data
- Grade Point Average (GPA) be used to measure student graduate rate
- Study period be used to measure student graduate rate
- Study program be used for mining process in order to find out a relationship between graduate rate with a study program

This data are formed of tables in one server. For mining process, graduate data and student master data has mixed with primary key NIM. After it had done the process of mining begins. The process of data integration performed when ETL (Extract, Transform and Load) process when built up a data warehouse on ETL process the data in data source combined into one in the data warehouse with key NIM.

Has searching the linkages between graduate rate with student master data. Student graduate rate can be seen from the long period of study and GPA (Grade Point Average). From that two parameters, data can be converted into a data type that allows for processing. Graduate rate can be measured based on long period of study and GPA, period of study for S1 had been categorized by the rule of ITB academical in 2013 point 6 declared that ITB organized education program who can finished by the student timely with normal ability, according to the curriculum. The students be expected can finished timely. Clause 6.1 period of study for bachelor's degree, normal period of study for bachelor's degree consist of (Anonymous, 2013):

- The first stage of the year is scheduled in 2 semesters or 1 year
- The undergraduate stage is scheduled in 6 semesters or 3 years

Whereas, GPA has categorized based on graduation predicate set out in SK Rektor Institut Teknologi Bandung about criteria for the predicate graduation bachelor's degree Alumnus which reads "Institute of Technology Bandung" give a graduation of bachelor's degree with predicate (graduation) as follows:

- Satisfy GPA with GPA 2.00-2.75
- Very satisfactory GPA with GPA 2.76-3.50
- Type GPA with the compliment with GPA 3.51-4.00

Categorization of graduate data based on the long period of study is Anonim:

- Appropriate to schedule, if the period of study not >10 semesters
- Not appropriate to schedule, if the period of study >12 semesters

From two categorization that we have talked about, Table 1 we can make a category based on combination of both data as follows.

Data analysis: Analysis process of mining to know the relationship between graduate data with the entry process data such as Table 2. USM is asmission test and SNMPTN is national selection entery higher education. From that early data be obtained the first candidate as Table 3. Set the threshold = 3, then the candidate who has the grade <3 will be erased. So that have a result as Table 4. From Table 3, we have the second candidate as

Table 1: Data information

Atribute	Description
A1	The long of study is 10 semester or <10 semesters and GPA 3.51-4.00
A2	The long of study is 10 semester or <10 semesters and GPA 2.76-3.50
A3	The long of study is 10 semester or <10 semesters and GPA 2.00-2.75
B1	The long period of study >10 semesters and GPA 3.51-4.00
B2	The long period of study >10 semesters and GPA 2.76-3.50
B3	The long period of study >10 semesters and GPA 2.00-2.75

Table 2: Preliminary data

NIM	Graduate category	Entry process
13207001	A1	USM
13307003	A2	SNMPTN
10108006	A1	USM
13308007	A3	SNMPTN
15009003	B2	SNMPTN
15507008	A3	SNMPTN
12008010	A3	SNMPTN
12108003	A2	USM
17507006	A2	USM
19009001	A2	USM
19009012	B2	SNMPTN

Table 3: First candidate

Itemset	Count
A1	2
A2	4
A3	3
B2	2
USM	5
SNMPTN	6

Table 4: The results after the threshold

Itemset	Count
A2	4
A3	3
USM	5
SNMPTN	6

Table 5: Second candidate

Itemset	Count
A2, USM	3
A2, SNMPTN	1
A3, USM	0
A3, SNMPTN	3

Table 6: The second result

Itemset	Count
A2, USM	3
A3, SNMPTN	3

Table 7: Preliminary data a second example

NIM	Graduate category	Entry process	Origin school
13207001	A1	USM	Bandung
13307003	A2	SNMPTN	Jakarta
10108006	A1	USM	Bogor
13308007	A3	SNMPTN	Bandung
15009003	B2	SNMPTN	Bandung
15507008	A3	SNMPTN	Jakarta
12008010	A3	SNMPTN	Bogor
12108003	A2	USM	Bandung
17507006	A2	USM	Garut
19009001	A2	USM	Bandung
19009012	B2	SNMPTN	Jakarta

Table 5. After the threshold obtained we can have result as in Table 6. From Table 6 can obtained the result as follows:

- Support A2, USM = Count (A2, USM)/number of transaction = 3/11
- Support A3, SNMPTN = Count (A3, SNMPTN)/ number of transaction = 3/11
- Confidence A2, USM = Count (A2, USM)/Count (A2) = 3/4
- Confidence A3, SNMPTN = Count (A3, SNMPTN)/ count (A3) = 3/3

Can be seen that mining process in the relationship between graduate rate and student entry process with threshold 3 produce a relationship A2, USM has a value support = 3/11 11, confidence = 3/4 and a relation A3, SNMPTN has a valued support = 3/11, confidence = 3/3. USM has a graduate rate A2 and SNMPTN has a graduate rate A3 so, it can be concluded that a student with the entry process through USM have a better graduate rate than a student with the entry process through SNMPTN.

Two examples from mining process to know a relationship between a graduate rate with the entry process and birth place data such as Table 7. From that early data we get the first candidate as in Table 8.

Table 8: First candidate, second example

Itemset	Count
A1	2
A2	4
A3	3
B2	2
USM	5
SNMPTN	6
Bandung	5
Garut	1
Jakarta	3
Bogor	2

Table 9: Results after the spacified threshold, the second example

Itemset	Count
A2	4
A3	3
USM	5
SNMPTN	6
Bandung	5
Jakarta	3

Table 10: Second candidate, the second example

Itemset	Count
A2, USM	3
A2, SNMPTN	1
A3, USM	0
A3, SNMPTN	3
A2, Bandung	2
A2, Jakarta	1
A3, Bandung	1
A3, Jakarta	1
USM, Bandung	3
USM, Jakarta	0
SNMPTN, Bandung	2
SNMPTN, Jakarta	3

Table 11: The second result, the second example

Itemset	Count
A2, USM	3
A3, SNMPTN	3
USM, Bandung	3
SNMPTN, Jakarta	3

Table 12: Third candidate

Itemset	Count
A2, USM, Bandung	3
A3, SNMPTN, Jakarta	3

Set the threshold = 3, then the candidate who have grade <3 will be erased. So that, have a result as Table 9. From the Table 3, we have the second candidate as in Table 10. After the threshold obtained we can have result as in Table 11. From the Table 11, we have the third candidate as in Table 12. From Table 12 above can obtained the result as follows:

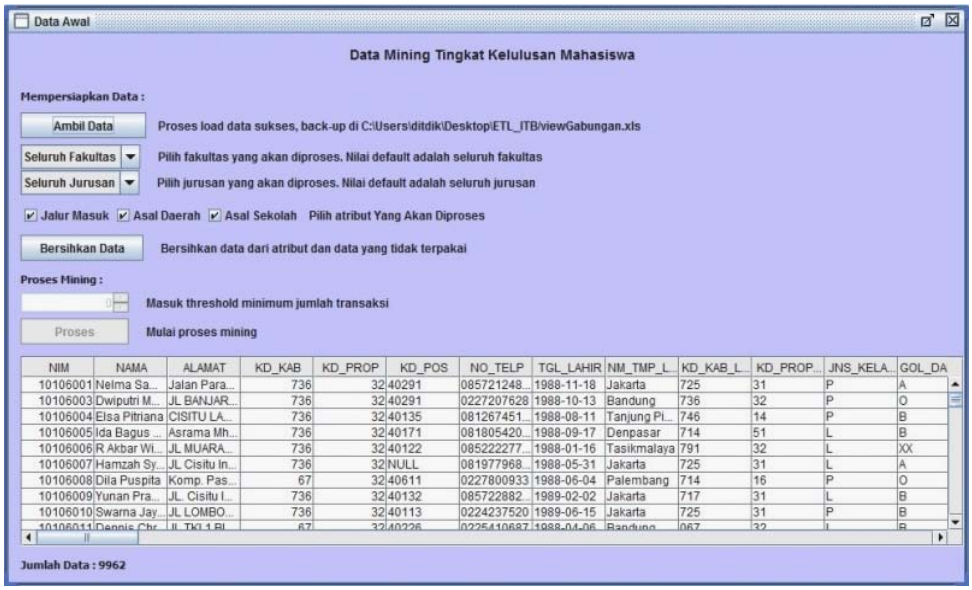

Fig. 1: Display design for first form in data mining application

- Support (A2, USM, Bandung) = Count(A2, USM, Bandung)/number of transaction = 3/11
- Support (A3, SNMPTN, Jakarta) = Count (A3, SNMPTN, Jakarta)/number of transaction = 3/11
- Confidence (A2, USM, Bandung) = Count (A2, USM, Bandung)/Count (2) = 3/4
- Confidence (A3, SNMPTN, Jakarta) = Count (A3, SNMPTN, Jakarta)/Count (A3) = 3/3

Be appointed that minimum support or threshold is 3. At the first iteration, the item that have a support or the countless than 3 must be eliminated from 1-itemset L1. And then 2-itemset C2 candidate from second iteration formed from cross product items which exist in L1. After 2-itemset candidate have counted from the database, 2-itemset 2 be appointed. The same process recur on the third iteration but besides {A2, USM, Bandung} and {A3, SNMPTN, Jakarta} who be candidate at the 3-itemset C3 actually there is also itemset {A2, USM, Jakarta} and {A3, SNMPTN, Bandung} which can be obtained from the combination of items at L2 but the two itemsets have to be cut because {USM, Jakarta} and {SNMPTN, Bandung} never been in L2. This process recurs until there is no new candidate which can be generated from the minimum threshold.

From the example of the Table 12 can be seen that algorithm apriori can reduce the number of candidates which is the support must be calculated by pruning. For example, a 3-itemset candidate can be reduced from 3. Reduction in the number of candidates is the main cause of the increase algorithm apriori performance.

Modelling and its output: In data mining application, there is two form. The first form is the first page which contains to order data retrieval, the selection of student master data attributes, threshold input, the instruction to mining process and exit button. The second form is a data mining report page which contains the result from data mining process that is support value table and confidence.

The orders on the first form is data capture button for the process of data collection, threshold input text to entering a threshold value, combo box input to choose a department, combo box input to choose attributes which will do a mining process, process button for order to mining process and exit button to exit the application. Besides the button, there is a result table from the order is given. Figure 1 and 2 display design at first form from the data mining application is Fig. 1.

The output of the data mining process presented in report form in data mining application. This form consists of two main information that is attributes information with the highest confidence for each graduate category and the table with confidence and support contain for each combination graduate rate and attributes. Besides, there is two order button that is back to the main menu button and exit from the application button. Display design for report form in data mining application as in the above.

Evaluation: In this data mining application testing used black box testing technique. Techniques used in black box testing among others (Pressman, 1997):

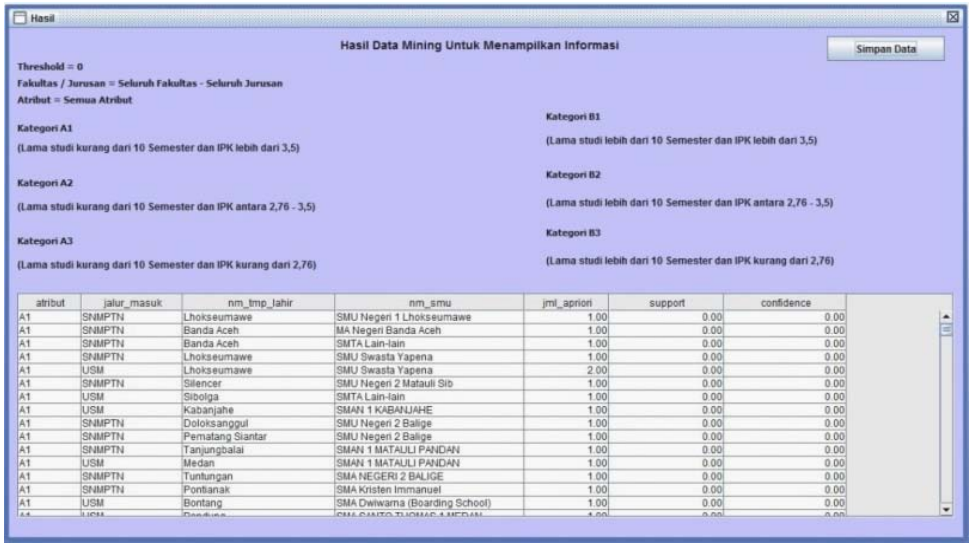

Fig. 2: Display design for report form in data mining application's

Table 13: Identification and implementation of testing

Test class	Test item	Test rate	Type of testing
Data retrieval function	Press the capture data button	Testing system	Black box
Data remove function	Press the remove data function	Testing system	Black box
Mining process function	Press the mining process button	Implementation testing	Black box

- Used for testing main functions from the designed software
- The righteousness of the software being tested just seen by output result from data or input condition given to existing function without seeing how the process to get the output and how the result from the mining process
- From the output, the program ability to fulfill the needs of the user can be measured at once can be seen the mistakes

Identification and the testing implementation can be seen in Table 13. A testing result considered a success if at the testing table, the result obtained in accordance with evaluation result criteria and expected results. Table test result can be seen in Table 14.

Deployment: Data mining application can be used to show graduate rate information. The show-up information in the form support value and confidence relationship between graduate rate and student master data. The higher the confidence value and support the stronger the value of the relationship between attributes. Student master data who mining processed such as entry process data, origin school data, student hometown and study program data. The result of data mining process can be used as consideration to make a next decision about factors that affect graduation rates, especially at student master data factors.

Research constraints: This research restricted at the information presentment about graduate rate with data mining technique. The presentment information in the form a support value and confidence relationship between graduate rate with student master data.

In this study, not discussed at the decision supporting system and academical information. In the develop data mining needs a warehouse data, therefore in this study discussed to develop a simple warehouse data who built up to fulfill the needs from data mining process. Data warehouse who built up not a warehouse data which save all the transactional data, only a data warehouse which support the development of data mining, so that the data and the format was adapted to the needs of data mining (Bramer, 2007; Witten and Frank, 2005).

This discussion is also limited at how to produce application who apply data mining techniques to produce information about the relationship between graduate rate wit student master data. In this study, not discuss at the data mining process result and output analysis result (Bramer, 2007; Preprocessing, 2006; Han *et al*., 2006). The discussion only at bachelor's degree program (S1) regular in general. Captured data is the student for bachelor's degree regular program at ITB. Student master data is attributed attached at student such as name, NIM

Table 14: Test result data mining application

Descripition	Prosuder testing	Expected output	Evaluation result criteria	The result obtained	Conclusion
Capture data function	Press the capture data button	The combined data view on table	At the combination gridview data combination show up	Data combination showed up in gridview combination	Accepted
Remove data fumction	Press the remove data button	The combined data view that is clean from a dirty data and not used attributes	Combination data that has showed up at the gridview clean combination	The clean data combination showed up at clean combination gridview	Accepted
Mining process function	Press the mining process button	Show report form's mining in the form of mining process result tables with support value and confidence and show each categories with the highest confidence value	Show report form with the result of mining process	Report form with mining process result table and the view for each categories with the highest confidence value	Accepted

(Student master number), address, origin school, etc. The graduate rate measured by the long period of study and IP. At this discussion, the long period of study and GPA refer to Academical rule 2013 number: 169/SK/I1.A/PP/2012 date 9th July 2012. Study period categorized based on Academical Rule clause 6.1 whereas IP categorized based on graduate predicate who set out in academical rule clause 5.5.

CONCLUSION

For the furthermore development of data mining application can be used another algorithm, for example is FP-Growth algorithm. The difference is apriori algorithm must do a scan database each time iteration whereas FP-Growth algorithm only does one-time scan database at the beginning. Other than that can be used to analyze the other data such as student Drop Out (DO) factors, TPB student graduate data, etc.

RECOMMENDATIONS

And this application can be developed by online, this means combined with the academical information system which has existed so that can facilitate access for the needy, especially for the leaders and analysis user.

REFERENCES

Anonymous, 2013. Regulation of academic. University of Makassar, Badan Penerbit UNM, Makassar, Indonesia.

Bramer, M., 2007. Principles of Data Mining. 1st Edn., Springer, London, ISBN: 978-1-84628-765-7, Pages: 344.

Han, J., M. Kaufmann and M. Kamber, 2006. Data Mining Concepts and Techniques. 2nd Edn., Elsevier Science & Technology, Amsterdam, Netherlands, ISBN: 9780123739056, Pages: 800.

Preprocessing, W.D., 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kauffman, San Francisco, California.

Pressman, R.S., 1997. Software Engineering: A Practitioner's Approach. The McGraw-Hill Companies, Inc., New York, USA., Pages: 4460.

Witten, I.H. and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edn., Morgan Kaufmann, San Francisco, California, ISBN: 0-12-088407-0, Pages: 525.