MINING ASSOCIATION RULES FOR EDUCATIONAL DATA

U Yin Moe Win

Lecturer,University of Computer Studies, Kalay

ABSTRACT

Educational data mining is a fast developing field of computer science and technology, which are helpful to enable end users for decision making process. One of the most important data mining applications is that of mining association rules. With the widespread use of medical information systems that include databases which consist of students records, course grade and name of subjects. The purpose of this paper is to develop a educational management system that can provide the possible decision and prediction information for students without the aids of knowledge expert. Finding frequent patterns plays an essential role in mining associations, correlations and many other interesting relationships among data. In this paper we analyzed the frequent pattern in education courses learning behaviour and discover the rules for association of these by using Apriori Algorithm.
.

Keyword : - *Association rule, Data mining, Apriori algorithm, educational datamining*

1. INTRODUCTION

Data Mining is the discovery of hidden information found in databases and can be viewed as a step in the knowledge discovery process [4, 5]. Data mining functions include clustering, classification, prediction, and link analysis (associations). One of the most important data mining applications is that of mining association rules. Association rules, first introduced in 1993 [2], are used to identify relationships among a set of items in a database. Association rules are one of the most researched areas of data mining and have recently received much attention from the database community [1].

Association Rule Mining finds interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored, a new research subject arise how can we find interesting association relations out of a large quantity of business transaction records to help make commercial decisions such as catalogue design, cross-marketing and loss-leader. A typical example of Association Rule Mining is market basket analysis. This process analyzes the purchasing habit of customers by drawing out relations between different commodities that are put into customers' baskets. Having learned which commodities are frequently bought at the same time by customers, retailers will be in a better position to make sales strategies. A well-known case of applying association model to practice is Beer and Diaper. By mining purchasing information of customers, supermarkets found a very useful rule: Among all the fathers that had bought baby diapers, 30% -40% also bought some beer. Subsequently, they changed the arrangement of shelves and placed diapers together with beer. As a result, the sales value increased substantively.

Apriori is the most classic algorithm of association rule mining. Apriori is an influential program for mining frequent itemsets for Boolean association rules. In our application, association of diseases set can be obtained by finding the frequent symptoms. Case (X, Symptom) => Cause (X, Diseases).

For example:

(1) fever, headache, malaise, nausea, vomiting, muscle and joint and eyeball pain => Influenza.

(2) fever with chills and rigors => Malaria. Symptom with one or more diseases can be obtained by finding Association Mining.

2. RELATED WORK

The issue of mining association rules between itemsets within customer transaction database is first raised in [1]. Agrawal et al. [3] presented the two new algorithms for solving the problem of discovering association rules between item in a large database of sales transactions, that are fundamentally different from the know algorithm. Apriori is the most classic algorithm of association rule mining. However, Apriori must be re-run to get new frequent itemsets when the minimum support changes, which will waste the latest mining results. The problem of mining association rules over basket data was introduced in [3].In this paper, we consider the problem of finding mining association rule by using Apriori algorithm.

3. MINING ASSOCIATION RULE

Association rule mining searches for interesting relationships among items in a given data set.
Association rule mining is a two-step process:

(1) Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a pre-determined minimum support count;

(2) Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence.

There are several properties of association rules that can be calculated.

Support: Support of a rule is a measure of how frequently the items involved in it occur together.

$$support(A \Box B) = P(A \Box \Box B) \quad (1)$$

Confidence: Confidence of a rule is the conditional probability of B given A.

$$confidence(A \Box B) = P(B \backslash A) \quad (2)$$

The statistical measures can be used to rank the rules and hence the predictions.

3.1 Apriori Algorithm

With the quick growth in e-commerce applications, there is an accumulation vast quantity of data in months not in years. Data Mining, also known as Knowledge Discovery in Databases(KDD), to find anomalies, correlations, patterns, and trends to predict outcomes.

Apriori algorithm is a classical algorithm in data mining. It is used for mining frequent itemsets and relevant association rules. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store.

It is very important for effective Market Basket Analysis and it helps the customers in purchasing their items with more ease which increases the sales of the markets. It has also been used in the field of healthcare for the detection of adverse drug reactions. It produces association rules that indicates what all combinations of medications and patient characteristics lead to ADRs.

3.2 Association Rules

Association rules [3] describe co–occurrence of events, and can be regarded as probabilistic rules. A good example of association rules is taken from the domain of sale transactions: an association rule in this domain expresses what items are usually bought together, information that is used for developing successful marketing strategies.

Association rules are if-then statements that help to show the probability of relationships between data items within large data sets in various types of databases. Association rule mining has a number of applications and is widely used to help discover sales correlations in transactional data or in medical data sets.

The basic task of the association rules analysis is to derive a set of strong association rules in the form of ("A1∧A2∧…….Am⇒ B1∧B2∧…….Bn") where Ai and Bj are sets of attribute values, from relevant data sets in a database.

The task of association rules mining is essentially to discover strong association rules in large database. In general, the problem of mining association rules is decomposed into the following two steps:

a) Discover the large itemsets, i.e, the sets of itemsets that have transaction support above a predetermined minimum support s. Itemsets with minimum support are called frequent itemsets.
b) Use the large itemsets to generate the association rules for the database.

Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets.
To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called Apriori property which helps by reducing the search space.
Apriori Property : All non-empty subset of frequent itemset must be frequent. The key concept of Apriori algorithm is its anti-monotonicity of support measure.

Table -1 Notations Used in Apriori Algorithm

k-itemset	An itemset having k itemset
L_k	Set of large k-itemsets (those with minimum support).
C_k	Set of candidate k-itemsets (potentially large itemsets).

Apriori Association Algorithm One of the most notable and effective association algorithms is the Apriori algorithm.
(1) Related Key Words About Association Rules
(a) Item: different values in a transaction. A data set is a collection of transactions.
(b) Itemset: A set of items is an itemset. Every transaction is an itemset. If an itemset X contains k items, it is called kitemset.Given a non-empty itemset X, a transaction T in dataset D contains X if X Í T. Maximum itemset is an itemset which consists of all the items, generally denoted by "I"[2].
(c) Association rules: An association rule is an induction rule of the dataset. It can be represented as $X \Longrightarrow Y$, where $X \subset I$, $Y \subset I$, and X and Y both are not empty, $X \cap Y = \emptyset$.Generally, the following two conditions are used to generate Association rules [2]:
1. Support condition: In transaction dataset D, where X and Y are non-empty, disjoint itemsets, the support condition of transaction $X \cup Y$ is $sup(X \cup Y) / N \geq \rho s$, ρs = support threshold. This formula means the ratio between the amount of
transactions $X \cup Y$ and the total amount N of transactions should be greater than support threshold.
2. Confidence condition: The confidence condition ensures that the mined rules are of high confidence. the rules should satisfy the confidence condition: $sup(X \cup Y) / sup(X) \geq pc$, ρc = confidence threshold. It means the probability that itemset Y occurs in a transaction where itemset X occurs, should be more than the confidence threshold.
(d) Frequent itemsets: If an itemset satisfies the support condition, it can be called a frequent itemset.
(2) Principle of Apriori Algorithm
Apriori algorithm was proposed independently by Agrawal Srikant, Mannila,Toivonen and Verkamo. Apriori algorithm can be divided into two steps:
1. To find out all frequent itemsets. Apriori algorithm utilizes a recursive mode to generate all frequent itemsets.
2. To find out all pairs of itemsets in step 1, which can be satisfy the confidence condition. Then, all the association rules can be acquired.
The pseudo code of the above two steps is produced below:
Ck: candidate k-itemset
Lk: frequent k-itemset
STEP 1:
(1) L1 = find out frequent 1-itemsets in database.
And record the amount of each item.
(2) for (k=2;Lk-1 $\neq \emptyset$;k++) {

(3) Ck can be generated by merging pairs of Lk-1.
(4) for each transaction t in database {
(5) for each candidate c ∈Ck
(6) if c is in t {c.count++}}
(7) prune those c∈Ck whose k-1 item-subsets
are not all in Lk-1
(8) Lk =those c in which c.count>=min_support }
(9) return L= all Lk;

STEP2:

(1) for (k=1;Lk ≠∅ ;k++)
(2) for (j=2;Lj ≠∅;j++)
(3) if x in Lj y in Lk and they satisfy the
confidence condition(>=min_conf)
(4) return x⇒y-x.

3.3 Data Pre-processing

The raw data which contains students' grade from Kalay Compute University. The data contains student ID, course and subject marks. All the observed students are from the same major. In total, there are 3000 lines of student information in the raw data, the five previous year data. All this data is stored in a CSV document.
In the raw data, some attributes such as gender and hometown are unnecessary, and some records are repetitive and redundant. Hence, the data needs to be cleaned up. And stored into MySQL After preprocessing, the clean data generated is shown in Figure 6.

	A	B	C	D	E	F
1	Student_i	Major	Course_id	Subjects_nam	Subject_s	Grade
2	10001	Computer	M	MYANMAR	56	B
3	10001	Computer	E	IELTS TEST BUI	67	B
4	10001	Computer	P	PHYSICS	80	A
5	10001	Computer	CST101	INTRODUCTIO	56	B
6	10001	Computer	CST102	MATHEMATIC	50	B
7	10001	Computer	CST103	COMPUTER AP	40	C
8	10001	Computer	CST104	PROGRAMMIN	41	C

Fig -2. The Clean Data

4. DATA MINING RESULT

Stored all candidate 1-itemsets and record the support count of each 1-itemset. The minimum support condition is set to 0.1 which means 10% of the total amount (the total amount is more than 300). The minimum confidence condition is set to 0.7 which is the conditional probability between two frequent itemsets. A part of C1 and L1 generated by Apriori algorithm is displayed in Table 2. and Table 3. respectively. Those candidates whose support counts are lower than the support condition will be pruned.

Table 2. Candidate 1-Items C1

Itemset	Amount
CST103:ICS:A	0
CST103: ICS:B	24

CST103: ICS:C	72
CST103: ICS:D	75
CST103: ICS:E	8
CST104:PL:A	0
CST104:PL:B	8
CST104:PL:C	42
CST104:PL:D	38
CST104:PL:E	14
……..	……..

Table 3. Frequent 1-Itemstes L1

Itemset	Amount
CST103: ICS:C	72
CST103: ICS:D	75
CST104:PL:C	42
CST104:PL:D	38
CST102:Mths:A	37
CST102:Mths:B	49
P:Phy:A	41
P:Phys:B	37
E:IELTS:B	59
E:IELTS:C	87
…….	…….

Table 4. Result of Association Rules

Rules	Confidence
CST102:Mths:C ⟹P:Phy:C	0.73
CS202:Mths:C⟹CS203:DS:C	0.8
CS202:Mths:C⟹203:DL:C	0.85
CST102:Mths:C⟹CST103:PL:C	0.73
CS206: SE:B⟹CS203:DS:C	0.81
CS204:SAD:B⟹CS201: JAVA:A	0.7
CS404:DBMS:B⟹CS405:UML: C	0.72
CS201:JAVA:A⟹CS203:DS:C	0.71
CS201:JAVA:A⟹CS202:Mths:C	0.71
CS304:UML:A⟹ CS305:CAT3:C	0.71
CS206:SE:B⟹CS204: SAD:B	0.76
CS304:DBMS:B:CS301:CO:C	0.7
..	……

Significant trends can be seen from the results of Apriori algorithm. For example, "CS204:SAD:B=>CS201:JAVA:A Confidence = 0.70 ", it can be said that if a student gets B level in SAD subject, he will excel in course CS201 , it means course CS204 may requires some similar skills as course CS201. In another word, students who are able to perform well in both course CS204 and CS201, certain parts of his/her learning skills are better than the rest of the students s' sample, vice versa. Another example, "304:DBMS:B =>301:APP:C Confidence=0.70", it tells us that a student gets C level in APP, he will probably get a bad level in course CS304:DBMS. That is very likely to say, there may be some similarity on the knowledge or skills requirement between course CS301 and course CS304. Students who cannot get a good level in both of the courses, he/she might be lacking of certain knowledge or learning skills, vice versa. According to above analysis, that can realize further.

5. CONCLUSIONS

The application of Educational Data Mining research is found out in educational field. The result also provides a good reference for education. In Future study is present a cloud-based framework to generate the useful rules via clustering algorithm.

The current research requires further study to discover more potential and valuable rules to get a clearer picture of the big data application in higher education and more importantly, to further optimize the education resource.

6. REFERENCES

[1] Abdullah Alshwaier , Ahmed Youssef and Ahmed Emam, "A Newtrend for E-Learning in Ksa Using Educationalclouds", Advanced Computing : an International Journal, 2012,Vol.3(1), p.81

[2] Agrawal, R.; Imieli_ski, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data – SIGMOD '93. p. 207.

[3] American Bar Association Section of Legal Education and Admissions to the Bar, Legal Education and Professional Development –An Educational Continuum Report of The Task Force on Law Schools and the Profession: Narrowing the Gap, July, 1992.

[4] Bael, S., S. H. Hat, and S. C. Parka. "Identifying gifted students and their learning paths using data mining techniques." Data Mining in ELearning(
Advances in Management Information) 4 (2006): 191-205.

[5] C Romero, S Ventura, Data mining in e-learning. WIT, 2006.

[6] David Chappell, (October 2008). " Introducing the Azure Services Platform An Early look at Windows Azure, .Net Services, SQL Services, And Live Services ". Chappell & Associates.157

[7] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996)."From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008.

[8] Jiaqu Yi, Sizhe Li, Maomao, Wu, H.H. Au Yeung Wilton W.T Fok, Ying Wang, Fang Liu "Apriori algorithm and K-Means Clustering algorithm based on Students' Information", 2014 IEEE Fourth International Conference on Big Data and Cloud Computing

[9] Luan, Jing. "Data mining and its applications in higher education." Newdirections for institutional research 2002.113 (2002): 17-36.

[10] MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering". Information Theory, Inference and Learning Algorithms. Cambridge University Press

[11] MacQueen, J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press.1-8.

[12] Merceron, A., Yacef, K. (2008). Interestingness Measures for Association Rules in Educational Data. In International Conference on Educational Data Mining, Montreal, Canada, 57-66.

[13] Minaei-bidgoli,B Tan, P., Punch, W. (2004). Mining interesting contrast rulesfor a web-based educational system. In International Conference on Machine Learning Applications, Los Angeles, USA,

[14] M.Lawanya Shri, Dr. S.Subha, "An Implementation of e-Learning System in Private Cloud", International Journal of Engineering and Technology, 2013, Vol.5(3), p.3036

[15] Paul Pocatilu. "Cloud Computing Benefits for E-learning Solutions".Oeconomics of Knowledge, 2010, Vol.2(1), p.9

[16] Peden, Elisabeth; Riley, Joellen, "Law Graduates Skills A Pilot Study into Employers Perspectives" [2005] LegEdRev 5; (2005) 15(1&2)Legal Education Review 87.
[17] Rakesh Agrawal and Ramakrishnan Srikant Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487- 499, Santiago, Chile, September 1994]
[18] Ramli, A.A. (2005). Web usage mining using apriori algorithm: UUM learning care portal case. In International conference on knowledge management, Malaysia, 1-19.
[19] Richard A. Huebner, "A survey of educational data-mining research",Research in Higher Education Journal, Retrieved 30 March 2014
[20] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D.Piatko, Ruth Silverman, Angela Y. Wu, "An Efficient k-MeansClustering Algorithm: Analysis and Implementation", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002
[21] "The NIST Definition of Cloud Computing". National Institute ofStandards and Technology. Retrieved 24 July 2011.
[22] University of Computer Studies , Kalay in Myanmar
[23] http://en.wikipedia.org/wiki/Microsoft_Azure
[24]http://azure.microsoft.com/en-us/documentation/articles/fundamentalsintroduction-to-azure/
[25] ttp://en.wikipedia.org/wiki/Platform_as_a_service
[26]http://en.wikipedia.org/wiki/IaaS#Infrastructure as_a_service_.28IaaS.29