Modelling Academic Resources: An Apriori Approach

Zunera Farooq
M.Tech Student
Department of CS & IT
University of Jammu

Vinod Sharma, PhD
Professor and Head
Department of CS & IT
University of Jammu

Muheet Ahmed Butt
Scientist D
PG Dept. of Computer Science
University of Kashmir

ABSTRACT
"Data Mining or Knowledge Discovery is the process of discovering patterns in large data sets" [1] in the form databases and data warehouses in structured or unstructured manner. Association Rule Mining (ARM) is primarily focused on analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most meaningful relationships. In the area of academic data mining, it concerns with developing methods for discovering knowledge from data that come from Academic Enterprise Domain. There are several data mining algorithms pertaining to association rules used both offline and online platforms. One of the most popular and classical is Apriori algorithm that is used to extract frequent itemsets from large database and generating the association rule for discovering the knowledge. In the proposed research we have implemented an Apriori Algorithm implementation using Matlab and Dot Net Technologies using an academic examination registration dataset. The various Association Rules have been used to mine valuable knowledge regarding present, past and future course selection trends on subjects selected by the students at undergraduate level. The results will provide an insight in making future decisions regarding proposing academic infrastructure pertaining to human resource development/management, building of new departments/ centers, enhancing/ reducing intake capacity for a course/subject etc in an optimized manner.

Keywords
Association rule mining, Support, Confidence, Lift, Apriori Algorithm.

1. INTRODUCTION
Data mining is considered a technique to discover the knowledge in the data and it is popular in big enterprises to generate patterns and trends out of that data. Data mining has been initially used for business purposes to find out customer interests, associations and trends pertaining to products. However, it has now been used in many other fields such as e-commerce management, business intelligence, research, and medical diagnostics where a large volume of data is handled as well as frequent patterns and trends are needed to be observed. Nevertheless, it has not been sufficiently applied in the academic sector, although there is an increasing amount of data accumulated in the educational institutions about students, courses, faculty, financial, content, administrative etc. Therefore, data mining can be very useful in this area because there are various associations of the data that can exist, in particular in the management of the educational system in an efficient manner. Therefore, it is currently an emerging interdisciplinary area where data can be used to deal with a wide scope to enhance the sector in an educational context. Even better, an underlying association with educational data can be revealed benefitting all the individuals who are anyways related to the system [15].

2. ASSOCIATION RULE MINING
Data mining has one technique known as "association rule mining" (ARM), ARM is the technique of finding interesting relations between the data elements in a large dataset [1]. Jiakang [2] illustrates that the purchase of a burger and French fry items yields the likeliness to buy drinks. This is the relationship of items that can occur frequently. Discovering this association is called association rule mining (ARM). The ARM involves finding of the item sets and generating association rules [3]. The occurrence of values in a database generates likeliness of other values. These values are referred to as items and these item sets are maintained in a transaction. The frequent itemsets are obtained based upon the support condition provided. In the ARM, it is important to understand several key terminologies used. The key terms are defined as follows [15]

2.1 Frequent itemset
A number of item sets appearing often by meeting the supporting threshold set, are the frequent item sets.

2.1.1 Support
Support is the number of transaction containing specific item set looking at whole transaction. Support(x) = no. of transactions containing x / Total number of transactions.

2.1.2 Confidence
Confidence is the percentage of transactions containing one item which comes along other item set in a database. In mathematical expressions, the confidence is shown as: Confidence(x -> y): support (x U y) / support(y).

2.1.3 Lift
The parameters Support and Confidence are used to determine if a rule is valid. However, there are times when both of these measures may be high, and yet still produce a rule that is not useful. Lift is defined as a measure to check whether the rule is in the list by random chance or we are expecting this association.

Lift = Confidence / Expected Confidence

2.2 Apriori Algorithm
The Apriori algorithm is a traditional data mining algorithm that is used to mine association rules from the given data [13]. The aim is to extract a set of strong association rules of the form X => Y i.e. items that satisfy condition X are most likely to satisfy Y also. Imagine we have a set of items I = {i1,i2,….,im}. Let D be a set of transactions where each transaction T be set of items such that $T \subseteq I$. A transaction T contains A i.e. a set of items if $A \subseteq T$. So, an association rule is of the form A => B where $A \subset I$, $B \subset I$ and $A \cap B = \emptyset$. The rule X => Y has support 's', in the transaction set D, if 's' percent of transactions in D contain X => Y. The confidence c means that c percent of transactions in D that contain X also contain Y. The steps involved in the Apriori algorithm are shown below [4]-13].

Ck: Candidate itemset of size k

Lk: frequent itemset of size k

L1= {frequent items};

for (k= 1; Lk!=∅; k++) do begin

Ck+1= candidates generated from Lk;

for each transaction tin database do: increment the count of all candidates in Ck+1that are contained in t

Lk+1= candidates in Ck+1with min_support

End

Return ∪kLk;

3. RELATED WORK

Association rule mining by Agrawal was first reported in [5]. Association rules are in the form of X => Y, where both X and Y are items or itemsets that are solely contained within a dataset and X ∩ Y = ø. This rule implies that X implies Y or that whenever X is present, Y will also be present. Since most datasets are very large, the user/client does not always want all the rules instead only those that are of interest/importance measures are needed so that the uninteresting rules can be removed.

Data Mining can be significantly used in educational field to enhance understanding of learning process in order to focus on identifying, extracting and evaluating variables related to the learning process of students [6] . Mining data in educational environment to look for interested hidden information is called Educational Data Mining.

Galit [7] in his case study use students data to analyze their learning behavior to predict the results and to warn students who are at risk of not perform before their final exams.

Pandey and Pal [8] researched the student performance by selecting 60 students from a college in india. The authors by means of association rule they find the interestingness of student in opting class teaching language.

Chandra and Nandhini [10], applied the association rule mining analysis based on students' failed courses to identify students' failure patterns. This is reported as a case study in [9].The aim of their study is to identify hidden relationships between the failed courses and suggest relevant roots of the failure to improve the performance of students. The mined association rules revealed some hidden patterns which could serve as a foundation for academic planners/thinkers in making academic decisions and aid in the curriculum re-structuring and modification with a view of boosting students' performance and reducing failure rate.

The authors of the research work in [6]-[11]-[12] presents the application of association rule mining on educational data to understand the knowledge and performance of students. The researchers implemented Apriori algorithm on student log data to bring out the interesting rules. Those rules are used to predict the students' performances and to impart the quality of education in the academic institutions. The algorithm using support measure generated frequent item sets in order to understand the interest of the students in the particular course. Interesting rules are produced based on frequent item sets on the dataset using confidence factor. The rule helps the teacher to understand the knowledge and performance of students in answering the questionnaire and understand the interest of the students in that course.

Educational data mining concerns with developing methods for discovering knowledge from data that come from educational domain. In paper [9] the authors used educational data mining to improve graduate students' performance and focussed on the problem of low grades of graduate students.

Suhem Parack, Zain Zahid et al in their paper have discussed the application of data mining in educational datasets for predicting academic trends and patterns. The authors modelled the approach by using apriori algorithm for student profiling which is one of the popular method for mining associations i.e. discovering co-relations among set of items [13].

The application of association rule mining in academic data mining is also reported by Hua Wand Ping Liu et al in their work Application of improved association rule algorithm in the courses management. The implementation of the algorithm is carried out in matlab [14].

4. PROPOSED MODEL

The proposed work is carried out in the following steps:

1. First of all data in a structured or unstructured format regarding under graduate students is extracted from a heterogeneous sources pertaining to many years and schemes and is consolidated in a single database structure. The structure of the data mostly include student rollno, registration number, name, parentage and subjects opted by the student. The ETL process is carried out using Dot Net Development Environment using Regex and ADO controls interface. Data Transformation phase is performed so that there is one to many relationship between the subjects and the roll number assigned to a student will be created.
2. Dataset of more than 30,000 student records is extracted from heterogeneous sources.
3. Once the Data is accumulated and transformed in a mat structure an Association rule Data Mining Apriori Algorithm is implemented using Matlab.
4. The results are converted back from mat structure after resolving various associations within the student enrollment and registration data, data is also analyzed to provide the knowledge regarding present, past and future course selection trends on subject group information.
5. A graphical knowledge modeling is also implemented in vb.Net using various charts to have an analytical overview of a prediction of student behavior.

5. APRIORI RESULTS

Table 1. Final Results Interpretation Arts Stream

S.no	Association Rules	Support	Confidence	Lift
1	ED-->GE	0.40794	1	1
2	HS-->GE	0.37474	1	1
3	PS-->GE	0.42622	1	1
4	UR-->GE	0.50013	1	1
5	[ED,UR]-->GE	0.29507	1	1
6	[HS,PS]-->GE	0.32134	1	1
7	HS-->PS	0.32134	0.85752	2.0119
8	HS-->[GE,PS]	0.32134	0.85752	2.0119
9	[GE,HS]-->PS	0.32134	0.85752	2.0119
10	PS-->HS	0.32134	0.75393	2.0119
11	PS-->[GE,HS]	0.32134	0.75393	2.0119
12	[GE,PS]-->HS	0.32134	0.75393	2.0119
13	ED-->UR	0.29507	0.72332	1.4463
14	ED-->[GE,UR]	0.29507	0.72332	1.4463
15	[ED,GE]-->UR	0.29507	0.72332	1.4463

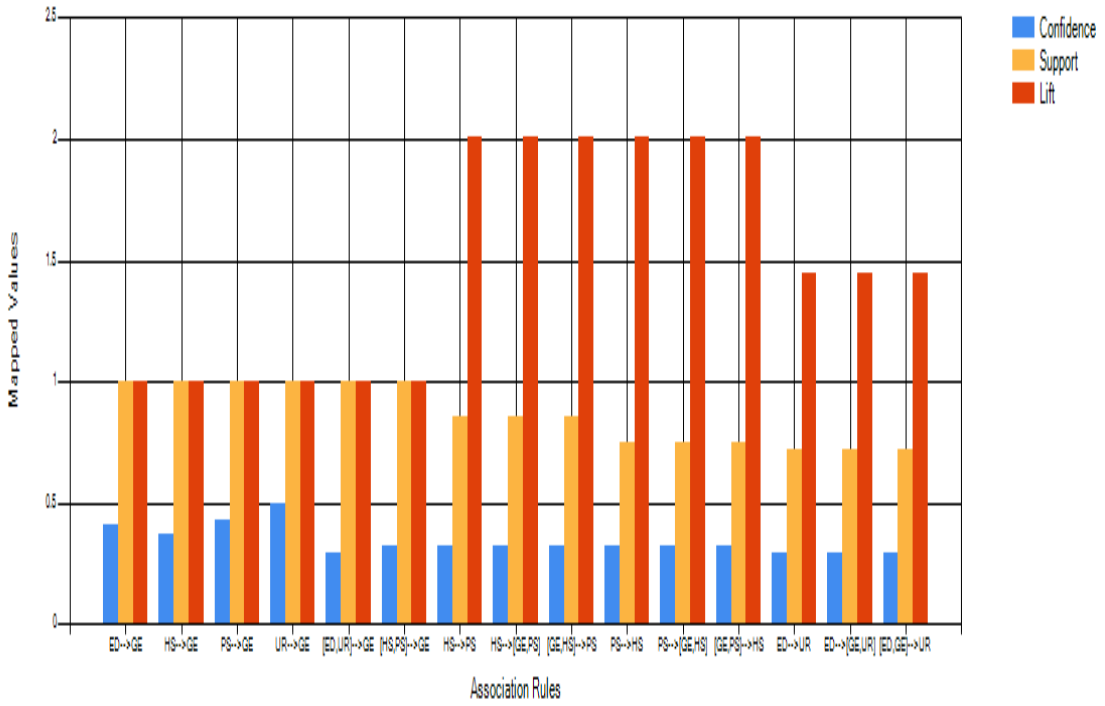

Fig 1: Chart depicting Apriori results for Arts stream

Table 2. Final Results Interpretation Medical Stream

S.no	Association Rules	Support	Confidence	Lift
1	[BO,CH]-->GE	0.61993	1	1
2	[BO,CH,ZO]-->GE	0.57108	1	1
3	ZO-->[CH,GE]	0.68002	0.93769	1.1245
4	BO-->[CH,GE]	0.61993	0.93769	1.1177
5	[BO,GE]-->CH	0.61993	0.93769	1.1177
6	[BO,ZO]-->[CH,GE]	0.57108	0.92667	1.1112
7	BO-->[CH,GE,ZO]	0.57108	0.85861	1.2626
8	[BO,GE]-->[CH,ZO]	0.57108	0.85861	1.2626
9	[CH,GE,ZO]-->BO	0.57108	0.8398	1.2626
10	ZO-->[BO,CH,GE]	0.57108	0.78747	1.2703
11	CH-->[BO,GE,ZO]	0.57108	0.68483	1.1112
12	[CH,GE]-->[BO,ZO]	0.57108	0.68483	1.1112
13	GE-->[BO,CH]	0.61993	0.61993	1

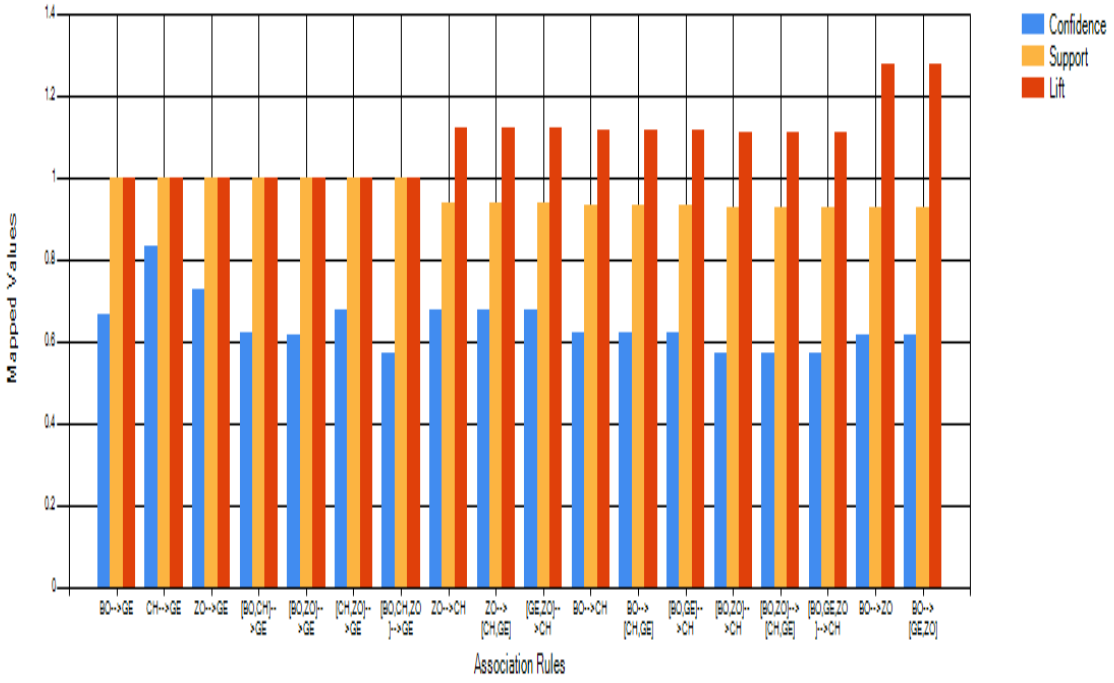

Fig 2: Chart depicting Apriori results for Medical stream

Table 3. Final Results Interpretation Commerce Stream

S.no	Association Rules	Support	Confidence	Lift
1	MAB-->[MD1,MD2]	0.52408	1	1
2	[MAB,MD1]-->MD2	0.52408	1	1
3	[MAB,MD2]-->MD1	0.52408	1	1
4	MAB-->[MD1,MD3]	0.52408	1	1
5	[MAB,MD1]-->MD3	0.52408	1	1
6	MAB-->[MD2,MD3]	0.52408	1	1

7	[MAB,MD2]-->MD3	0.52408	1	1
8	[MAB,MD3]-->MD2	0.52408	1	1
9	MAI-->[MD1,MD3]	0.47592	1	1
10	[MAI,MD3]-->MD1	0.47592	1	1
11	MAI-->[MD2,MD3]	0.47592	1	1
12	[MAI,MD2]-->MD3	0.47592	1	1
13	[MAI,MD3]-->MD2	0.47592	1	1
14	MD1-->[MD2,MD3]	1	1	1
15	MAB-->[MD1,MD2,MD3]	0.52408	1	1
16	[MAB,MD1]-->[MD2,MD3]	0.52408	1	1
17	[MAB,MD2]-->[MD1,MD3]	0.52408	1	1
18	[MAB,MD1,MD2]-->MD3	0.52408	1	1
19	[MAB,MD3]-->[MD1,MD2]	0.52408	1	1
20	[MAB,MD1,MD3]-->MD2	0.52408	1	1
21	[MAB,MD2,MD3]-->MD1	0.52408	1	1
22	MAI-->[MD1,MD2,MD3]	0.47592	1	1
23	[MAI,MD2]-->[MD1,MD3]	0.47592	1	1
24	[MAI,MD1,MD2]-->MD3	0.47592	1	1
25	[MAI,MD3]-->[MD1,MD2]	0.47592	1	1
26	[MAI,MD1,MD3]-->MD2	0.47592	1	1
27	[MAI,MD2,MD3]-->MD1	0.47592	1	1

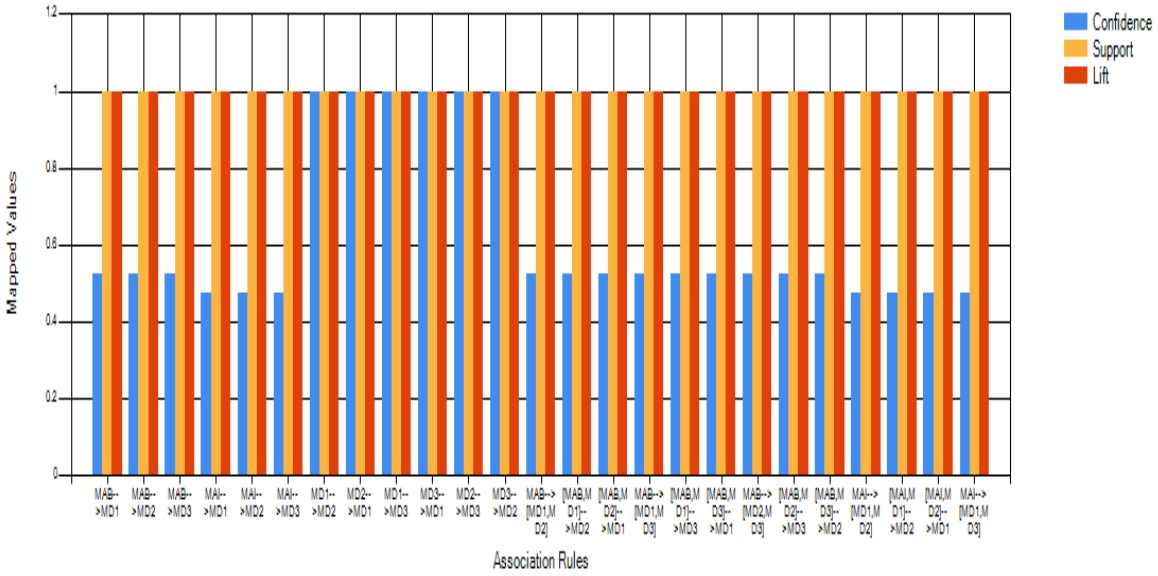

Fig 3: Chart depicting Apriori results for Commerce stream

6. OBSERVATIONS

The dataset that was used in the proposed research included 310 subject groups of around 78 subjects and around 30000 students at undergraduate level. The observation that were made by these results from the said dataset is as under:

1. There is more trend towards arts stream than science, home science and commerce streams.
2. It is also observed that the subjects like English, History, Education, Urdu and Political Sciences are taken by more students than other subjects.
3. Similarly the table I can depict the other association rules generated for Arts stream.

4. Table 2. Shows the rules generated in medical stream. The rules show association in subjects BO (Botany), ZO (Zoology), CH (Chemistry) etc.
5. Apriori algorithm was executed on commerce sub data set also to generate the association rules shown in Table 3.
6. Rest of the subjects have very low confidence values hence are not displayed in the final results.

7. CONCLUSION

Academic Mining is a field of study for data mining where the various techniques can be applied to find out the undisclosed relationships between the data items. This research work proposes an association rule based Apriori data mining technique on educational dataset for helping university authorities to manage the infrastructure and resources optimally while opening of new courses in the educational institutes by finding hidden relationships and associations from the data of more than 30000 record of students. A graphical knowledge modelling is implemented using various charts to have an analytical overview of a prediction of student behavior. The student record dataset can be tested with other association rule mining algorithms in future to generate association and relationships. Moreover, the student record can also be analysed to find out more hidden rules that can improve the prediction mechanism of academic mining for welfare of students and academic institutions.

8. ACKNOWLEDGMENTS

First, I would like to thank my Supervisor Professor Vinod Sharma for providing me the opportunity to work under him. I am so deeply grateful for his help, valuable guidance and support. I would like to thank all the faculty members, batch mates of Department of Computer Science and IT for helping me in any way. Many Thanks to Mr. Muheet Ahmed Butt and all the experts, authors whose research and publications were helpful in compiling this research paper.

9. REFERENCES

[1] Bhargava, M., & Selwal, A. (2013). Association Rule mining using Apriori Algorithm: A review. *International Journal of Advanced Research in Computer Science*, *4*(2).

[2] Jianking, N. (2013). Application of apriori algorithm to customer analysis. *Information Technology Journal*, *12*(6), 55-59.

[3] Oguz, D., Yildiz, B., & Ergenc, B. (2013). DMA: Matrix Based Dynamic Itemset Mining Algorithm. *International Journal of Data Warehousing and Mining (IJDWM)*, *9*(4), 62-75.

[4] Wang, P. J., Shi, L., Bai, J. N., & Zhao, Y. L. (2009, December). Mining association rules based on apriori algorithm and application. In *Computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on* (Vol. 1, pp. 141-143). IEEE.

[5] R. Agrawal, "Mining Association Rules between Sets of Items in Large Databases," ACM SIGMOD Int. Conf. Manag. Data, vol. 22, no. 02, pp. 207–216, 1993

[6] B. K. Baradwaj, "Mining Educational Data to Analyze Students ‟ Performance," Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 6, pp. 63–69, 2011.

[7] Galit.et.al, "Examining online learning processes based on log files analysis: a case study". Research, Reflection and Innovations in Integrating ICT in Education 2007.

[8] U. K. Pandey, and S. Pal, "A Data mining view on classroom teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.

[9] M. M. A. Tair and A. M. El-halees, "Mining Educational Data to Improve Students ' Performance : A Case Study," vol. 2, no. 2, pp. 140–146, 2012.

[10] Chandra, E. and Nandhini, K. (2010) 'Knowledge Mining from Student Data', European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163.

[11] S. V. Cristóbal Romero, "Educational Data Mining: A Review of the State of the Art," IEEE Trans., vol. 20, no. 10, pp. 1–19, 2010.

[12] S. Khadijah and Z. Tasir, "Educational data mining : A review," Procedia - Soc. Behav. Sci., vol. 97, pp. 320–324, 2013.

[13] Parack, S., Zahid, Z., & Merchant, F. (2012, January). Application of data mining in educational databases for predicting academic trends and patterns. In Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on (pp. 1-4). IEEE.

[14] Wang, H., Liu, P., & Li, H. (2014, June). Application of improved association rule algorithm in the courses management. In Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on (pp. 804-807). IEEE.

[15] Bhandari, P., Withana, C., Alsadoon, A., & Elchouemi, A. (2015, October). Enhanced Apriori Algorithm model in course suggestion system. InComputing and Communication (IEMCON), 2015 International Conference and Workshop on (pp. 1-5). IEEE.

[16] Mirajkar, A. M., Sankpal, A. P., Koli, P. S., Patil, R. A., & Pradnyavant, A. R. (2016). Data Mining Based Store Layout Architecture for Supermarket.