AN ARCHITECTURE FOR ASSOCIATION RULE DISCOVERY FROM STUDENTS REPOSITORY

[1]MOHAMMAD KAMRAN, [2]S. QAMAR ABBAS, [3]MOHAMMAD RIZWAN BAIG,

[1] Research Scholar, Integral University, Lucknow, India

[2] Professor, Ambalika Institute of Management & Technology, Lucknow, India

[3] Professor, Integral University, Lucknow, India

***Abstract*- The educational systems currently face number of issues such as high dropout rates, identifying students in need, personalization of training, predicting the quality of student interactions and placement of students. Data mining provides a set of techniques, which can help the educational system to overcome these issues. One of the biggest challenges that higher education faces today is predicting the academic paths of student. To better manage and serve the student population, the institutions need better assessment, analysis, and prediction tools to analyze and predict student related issues. These tools can be very helpful in managing and assisting students in higher education that serve thousands of students for job placement and higher studies. In this paper, we proposes data mining tool that can help to improve an education system by enabling better understanding of the students.**

***Keywords*— Data Mining, Association Rules, Apriori, Rule discovery for education, clustering.**

I. INTRODUCTION

Many higher education systems are unable to guide students in selecting career paths, deciding majors, and detecting student population who are likely to drop out because of lack of information and guidance from the school system. Data mining can help to improve an education system by enabling better understanding of the students. The information extracted through mining can help the teachers to manage their classes better and to provide proactive feedback to the students. The identified patterns using association rule mining technique are analysed to offer a helpful and constructive recommendations to the academic planners in higher institutions of learning to enhance their decision making process. This will aid in the curriculum structure and modification in order to improve students' academic performance and trim down failure rate.

Han and Kamber [2] define data mining as the process of discovering 'hidden images', patterns and knowledge within large amount of data and making predictions for outcomes or behaviors. Data mining methods can help bridge the knowledge gaps in higher educational system.

II. MOTIVATION

The data mining application in the area of education is wide spread. The researchers have explored various applications of data mining in education. The authors had gone through the survey of the literature to understand the importance of data mining in higher education. The research papers mostly concentrated on the data mining application from domain perspective. We had tried to analyze its importance from higher education perspective which has not been explored as much.

TABLE I
Data Mining Research Done Related To The Context Of Higher Education.

S. No.	Author	Year	Work
1	Ma et al	2000	Presented a real life application of data mining to find weak students
2	Luan J.	2001	Introduced a powerful decision support tool, data mining, in the context of knowledge management.
3	Luan J.	2002	Discussed the potential applications of data mining in higher education & explained how data mining saves resources while maximizing efficiency in academics.
4	Delavari et al	2005	Proposed a model for the application of data mining in higher education.
5	Shyamala, K. & Rajagopalan, S. P.	2006	Developed a model to find similar patterns from the data gathered and to make predication about students' performance.
6	Sargenti et al	2006	Explored the development of a model which allows for diffusion of knowledge within a small business university
7	Ranjan, J.	2008	Examined the effect of information technology in academic institutions for sharing information

First, an educational institution often has many diverse and varied sources of information. There are the traditional databases (e.g. students' information, teachers' information, class and schedule information, alumni information), online information (online web pages and course content pages) and more recently, multimedia databases. Second, there are many diverse interest groups in the educational domain that give rise to many interesting mining requirements. For example, the administrators may wish to find out information such as the admission requirements and to predict the class enrollment size for timetabling. The students may wish to know how best to select courses based on prediction of how well they will perform in the courses selected. The alumni office may need to know how best to perform target mailing so as to achieve

the best effort in reaching out to those alumni that are likely to respond. All these applications not only contribute an educational institute in delivering a better quality education experience, but also aid the institution in running its administrative tasks effectively. With so much information and so many diverse needs, it is foreseeable that an integrated data mining system that is able to cater to the special needs of an educational institution will be in great demand. The school management wanted to answer the following questions

1. How can we predict the job absorption rate for a particular branch i.e. to figure out the percentage of students that will be placed in a particular branch in a particular year in the future?
2. How fast can we understand the performance of each student related to different attributes?
3. How can we ensure quality of delivery of education to every child while teachers keep on changing?
4. How can we convert information about individual students into actionable knowledge?
5. How can IT be used for answering these questions as has been used in the corporate and scientific communities?
6. How can these analysis be done in minutes as compared to weeks and days?

III. METHODOLOGY

Not all that shines is gold, but, some things that do not always shine are precious as gold, and so it is with the minds of our children. The meaningful knowledge and potentially useful patterns extracted through data mining can assist in improving the quality of education and performance of students. We describe the conceptual frame work of data mining process in student education. The framework helps institutes to explore the effects of probable changes in recruitments, admissions and courses and ensures efficiency in the quality of students, student assessments, evaluations and allocations. This study formulate an interface that uses a database of institute and predicts methodology by which 100% placement can be achieved for a particular branch in a particular year and would also come up as aid to the students which can guarantee them a secure future.

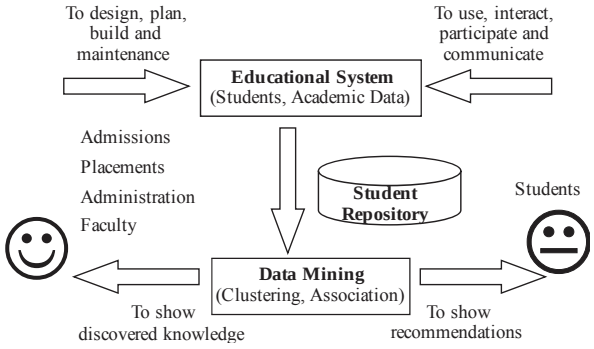

Fig. 1 Architecture of a Data Mining Model

Even students who are often bypassed by traditional recruiting activities have been known to demonstrate such potential. Potential often unrealized, wasted, lost. Those who evaluate students make their future. How we evaluate them decides the future of our nation. We have a national responsibility to ensure that every student is correctly assessed for all the potential that is our nation's to harness and nurture. Assessment practices, more than any other education responsibility, require our national attention. Emerging data mining technologies will attempt to do just that. But that is not enough. Today we are faced with the additional challenge of not indulging in the luxury of bypassing any human potential. We must apply those technologies where they will make the most difference, to mine the potential of our youth.

There is an increasing and compelling national need for making identification of academic promise to enhance the quality and quantity of the work force by drawing qualified candidates for engineering and high-tech careers from our nation's entire population of students.

In a technology and information driven society, a preferable approach is to combine advanced technology with methods for alternative assessment to identify human potential.

Discovering student strengths and weaknesses is not part of standard achievement testing practices, rather assessing current performance is. What can be learned about students by the use of data mining techniques, however, is highly subjective to the data being collected. Typically, assessment practices do not elicit data that can be useful in the discovery of new knowledge about student potential. Data mining allows for the generation of specific and general queries (about student performance), and these are possible if appropriate data has been collected and is present in the assessment process. The discovery of new knowledge about performance, using data mining, may be able to identify highly qualified individuals who are missed by standard assessment practices, if the necessary data is collected.

The methodology consists mainly of six steps as indicated in figure 1: Collecting the relevant features of the problem under study, preparing the data, clustering and building the association model, evaluating the model using one of the evaluation methods, and finally using the model for future prediction of the student performance. These steps are presented in the next subsections.

IV. PROPOSED WORK

Our approach for data mining is a hybrid approach that consists of two data mining techniques namely: Clustering and Association Mining. The clustering technique is applied on the composite records and valid episodes of events records from the replicates of operational database entities in the data warehouse to partition the generalized composite records based on the pattern of valid episode of events. The association mining phase works on the appropriate partition to generate rules related to the requested mining task. The main contribution of this work is the presentation of a new technique in the mining phase. This phase is based on integrating clustering partition and discovery of association

rules among different attributes in a data warehouse. Nowadays, there are many techniques to obtain association rules. These techniques can be shown on the well-known Apriori algorithm [11], [3], [4], [10], [8], [9] and many extensions shown in [5],[6],[12],[13],[7]. These techniques are modified to suite dynamic educational data characteristics.

The proposed technique is to perform association mining based on prior domain knowledge. The proposed system consists of two main phases: preprocessing and building a data warehouse phase, and data mining phase. These phases are interconnected through a repository management system [7] that contains an ontology for students data, data warehouse definitions, and preprocessing and transformation algorithms definitions

Data Mining Phase

Students' data quality may be insufficient if data is collected without any specific analysis in mind [14]. In Educational institutions there are many diverse yet interesting databases available ranging from students, faculty, courses, admin to research and consultancy, infrastructure facilities etc. It is designed to keep track of students' data, credentials and academic results to stream line the process of students' performance.

Data Collection and preparation

We collect the data from student database. To get better input data for data mining techniques, we did some preprocessing for the collected data. After we integrated the data into one file, to increase interpretation and comprehensibility, we discretized the numerical attributes to categorical ones. For example, we grouped all grades into three groups Excellent, Pass & Failure as described in table II below. In the same way, we discretized other attributes such as attendance, 10+2 grade, activities, percentage of practical session, exercise given by teacher, final mark etc. as shown in Table III.

TABLE II
VALUES OF FINAL MARK

Final percentage	Description	Possible Values
X < 35%	Fail	C
35% ≤ X ≤ 70%	Pass	B
X > 70%	Excellent	A

Collecting the Relevant Features

For this step, the collected data were prepared in tables in a format that it is suitable for the used data mining system. Initially more than 20 attributes have been collected and some of the attributes have been manually eliminated since they are considered as irrelevant to the study. The data are cleansed by removing the various inconsistent values using the same standard value for all the data. The cleaning also includes filling out the missing values using the most majority data approach. Finally, the most significant attributes list contains the following attributes presented in descending order according to their ranks:

TABLE III

Feature	Alias	Description	Possible Values
Enrolment No.	ENR	Enrollment number at the time of admission	Yes, No
Attendance	ATT	Attendance in total semester	A, B ,C
Age of the student	AGE	Age of the student.	18,19,20,21,22
10+2 Grade	INT	Intermediate result	A, B, C
Area of expertise	EXP	Area of expertise of the student (mathematics, computer, electronics etc.).	M,C,E
Gender	G	Student's gender.	M, F
Fund	F	Financial position of students (Private, Scholarship, Finance)	P, S, F
Student Department	STD	Branch of the student	ME, CS, IT
Activities performed by the student	ACT	Percentage of the activities performed by the student with respect to the total activities of the course.	A, B, C
Percentage of practical session	PSA	Percentage of student's practical and project session.	A, B, C
Exercise given by teacher	ET	Assignment's to be completed at home	A, B, C
Average mark of the experience report	ER	Average mark obtained by the student in the experience report. In this report the student evaluates his/her learning process and describes the main concepts learned.	A, B, C
Final mark	MARK	Final mark obtained by the student in the course.	A, B, C
Evaluation	EVL	Final Evaluation of students	A, B, C

A=Excellent, B=Good, C=Poor

V. CLUSTERING

Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group [2]. In educational data mining, clustering has been used to group students according to their attendance. The clustering process consists of two main tasks. The first task is to generate generalized composite records by dropping specific information. The second task is to cluster the generalized composite records according to the effect of the episode of events.

In case that new transactions are added to the operational database, the data warehouse composite record will be appended accordingly. The clustering algorithm is designed to maintain the rules generated incrementally, i.e. not to repeat the clustering procedure from scratch; a process that takes

long time. The incremental updating method takes into consideration the occurrence of one of the following events: A composite record is moved from one group (generalized composite record) to another in the same cluster, which affects weights of both groups within the cluster. A composite record is moved from one cluster to another. This will affect the number of the generalized composite records in the two updated clusters. A new generalized composite record is added to a cluster, which increase the number of records in this cluster by the weight of the generalized record. An existing group is deleted from a cluster, which decrease the number of generalized records within the cluster by the weight of this record. There is also a possibility that a complete new cluster is created and hence all needed adjustments will be done.

```
\* Create generalized composite record*
for each composite record do
{
Drop specific information which is enrolment no.,
attendance from composite records.
If this record exist in the generalized composite table
     Then add 1 to -weight field,
          Else create a new record in the generalized
composite table
}
for each generalized composite record do
 {
If the episode of events not exist into valid episode of
events table (table 1),
     then if the logical sequence of historical transactions
     of generalized- composite record is valid
               then , add the new episode of events to valid
               episode of event table,
     else write the erroneous code of event into invalid
     episode of event partition to be examined by the expert
     user.
else get the effect on portfolio from valid episode of events
table and write the generalized record in the appropriate
cluster table
}
```

Clustering process algorithm

VI. ASSOCIATION RULE ALGORITHM

Association rule mining is one of the most well studied mining methods. Such rules associate one or more attributes of a dataset with another attribute, producing an if–then statement concerning attribute values. Mining association rules between sets of items in large databases was first stated by Agrawal, Imielinski, and Swami [15] and it opened a brand new family of algorithms. The original problem was the market basket analysis that tried to find all the interesting relations between the bought products. Sequential pattern mining [16] attempts to find inter-session patterns such as the presence of a set of items followed by another item in a time-ordered set of sessions or episodes. These methods have been applied to education systems. Association rules are very useful in Educational Data Mining since they extract associations between educational items and present the results in an intuitive form to the teachers. Furthermore, they require less extensive expertise in Data Mining than other methods. Association can be used to track students activities related to placement, discipline programs, attendance regularity, specializations and courses.

The algorithm is based on well known existing techniques to obtain association rules as Apriori algorithm. This algorithm is modified to enable a user to control and impose his area of focus during knowledge discovery steps in order to overcome the loss of information problem and to enable him/her to generate rules that he/she is interested in. Loss of information problem occurs as result of discretization. The proposed algorithm solved this problem by allowing the user to define the relative weight or support of each attribute interval category such that the mining algorithm could generate rules using this attribute interval category only if this support is satisfied. For example if the total number of composite records is 100 in a certain cluster and the user specifies the support as 10%, for a certain interval category of an attribute, rules can only be generated from the generalized composite records that contain this attribute category interval, if the total weight of these records is at least 10.

This technique will enable the user to choose the relevant attribute value by giving it a small relative weight to enforce the mining algorithm to generate rules having this attribute value in their premises even if the number of occurrence of records having this attribute value is small. Less relevant attribute value is to be given higher support such that it will be pruned if the number of occurrence of records having this attribute value is small.

1- Select the mining task and consequently the appropriate cluster
2- Get the confidence threshold for generating a rule (this means that the rule will only be generated if the number of occurrences of records described by this rule divided by the total number of records in the cluster greater than the given confidence threshold)
3- Construct a matrix (calculated relative weight) with number of rows equal to the number of attributes (m) and number of columns (n) equal to the maximum number of categories of a certain attribute
4- Using the appropriate cluster, fill in the calculated relative matrix with the relative weight of each attribute category in this cluster
5- Compare the calculated relative weight with the user given support and mark irrelevant attributes categories.
6- For each generalized composite record do
7- For each generalized composite record attribute do { if the attribute category is irrelevant then mark it as irrelevant copy relevant attributes category into a new table}
8- Group similar rows in the new table and calculate a confidence value for this grouped record
9- Generate rules

Proposed algorithm

The algorithm uses the data stored in each cluster to generate rules that describe student's performance and evaluation. Mining each category of these rules is defined as a mining task. The generated rule premises are a subset of instance of factors affecting portfolio status and the conclusion part is an episode of events that cause this effect.

VII. SYSTEM ARCHITECTURE

The system architecture is shown in Figure 2. The database resides in the server machine. The stored procedures (Oracle) reside in the server side. Our VB application runs in the client machine. It consists of several modules: LogIn, Rule Generator, Import Data and Visualization module. LogIn module is used to connect to the database server. Rule Generator is used to mining the association rules given the information provided by the user. These modules can be accessed using the Main window.

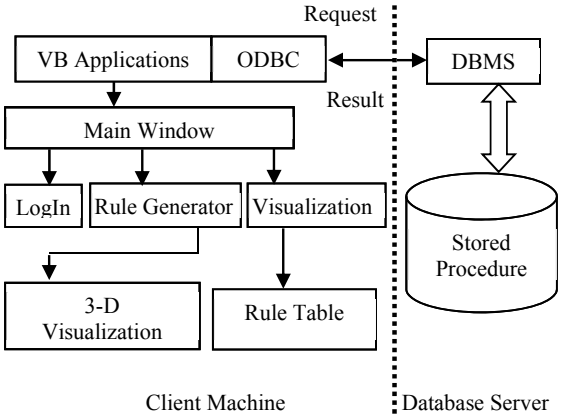

Fig. 2 System Architecture

VIII. IMPLEMENTATION METHODOLOGY

At the beginning of any mining task, the system acquires the support for each attribute category defined at discretization step during preprocessing phase of a generalized composite record in the corresponding cluster. Figure 3 depicts the user interface screen that acquires these supports. In order to show how our technique has enhanced the rule generated, we conducted the following experiment steps: Run the system and give variable support for each attribute category based on the user interest.

1) Count the number of rules generated and the number of used premises in these rules.
2) Rerun the system and give equal support for all attributes categories.
3) Count the number of rules and the premises used in these rules.
4) Examine the quality of rules generated in each case by comparing the number of rules and premises used.

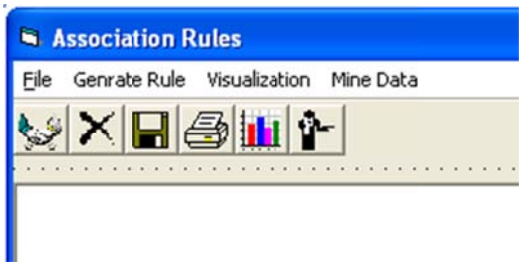

Fig. 3 User Interface Window of Mining Tool

The main window consists of the menu, toolbar, and a text area. The operations of rule generation and rule visualization are mainly done through the menu. Under the "File" menu, there are several submenus, such as "Connect", "Disconnect", "Save to File", "Clear Messages", "Exit". The submenu "Connect" is used to let the user connect to a specified DBMS, such as Oracle, since this software is designed to be able to retrieve the data from Oracle. "Disconnect" submenu is to simply let the user disconnect from the DBMS after he finished the mining operations. But if the user forgot to disconnect from the DBMS, when the application is terminated, the connection will be disconnected automatically. "Save to File" and "Clear Messages" submenus are related to the log messages generated by the mining operations. For each association rule generation algorithm and each input data set, there will be the some messages output to let the user know what had happened. After each operation, the user can choose to save these messages into a text file or clear these messages from the window. For each major function, such as "Connect", "Disconnect", "Rule Generation", "Exit", there is an icon on the toolbar corresponding to the operation for the user's convenience. So the user can just click the icon on the toolbar once instead of using the menu.

IX. EXPERIMENTAL ANALYSIS

TABLE IV Generated Rules

Rule #	Rules	# STD	# Attrib
7	IF ENR = Y, ATT = A, INT=A, G = M, STD=IT, ACT=A, PSA=A, ET=A, ER=B, MARK=A THEN EVL = A	13	10
3	IF ENR = Y, ATT = B, INT=A, G = F, STD=IT, ACT=A, PSA=A, ET=A, ER=B, MARK=A THEN EVL = A	9	10
11	IF ENR = Y, ATT = B, INT=A, G = M, STD=CS, ACT=A, PSA=C, ET=C, ER=B, MARK=B THEN EVL = B	9	10
17	IF ENR = Y, ATT = C, INT=A, G = M, F=SC, STD=ME, ACT=A, PSA=B, MARK=B THEN EVL = A	8	9
9	IF ENR = Y, ATT = A, INT=B, G = F, ACT=A, PSA=A, ET=A, ER=B, MARK=B THEN EVL = B	5	8
14	IF ENR = Y, ATT = C, G = M, ACT=B, PSA=A, ER=B, MARK=A THEN EVL = A	4	7
3	IF ENR = Y, ATT = C, MARK=C THEN EVL = C	3	3

How Generated Rules Could Serve Strategic and Tactic Decisions

Association rules support both strategic and tactical decision making for each mining task in the system. An example of strategic decision is to decide what types of students are eligible to quality for final placements as per company's requirements, which student is specialized in what and the regularity of the students. Or the student can know

the best course, best program based on prediction of how will they perform in the courses selected. The analysis addresses these types of issues and aids in contributing the increase of quality education delivery. It will also address issues like what types of students are eligible to quality for final placements as per company's requirements. The teacher can also concentrate on which subject students is weak and also if the students has to recommended for further studies the appropriate counseling can be done.

The benefit of this method is that it can predict low grades students on time. For example the teacher can predict weak students before the end of the semester and he may work on them to improve their performance before the final.

Some of the generated rules are given in Table IV in a form that is understandable by humans. In Table IV, the first column represents the rule number, the generated rules are presented in the second column, the number of the students who successfully satisfy the rules is given in the third column, and the number of attributes contained in the rule is given in the last column. The table shows the rules in a descending order depending on the number of the students who successfully have satisfied the rule. This ordering helps in determining the most significant rule. For the generated rules, the longest rule consists of 10 attributes while the shorter rule contained only 3 attributes.

X. CONCLUSION

We have presented a new rule extraction method based on fast and effective association rule mining algorithms, to help in enhancing the quality of the higher educational system by evaluating student attributes that may affect the student performance. This method has been used on a data set obtained from a educational institute to analyze students different attributes. The experimental results have shown that the extraction rules method presented in this paper was able to obtain comprehensible, actionable and realistic logical rules describing students' different attributes. The research is also planned for testing suitability of these techniques in appraisal of teachers, analyzing activities of students and expectations of parents, alumni interaction etc.

This knowledge could be used for real time student personalization guidance, and to help teachers in enhancing the students' performance. For this knowledge to have an intuitive and useful form, results have been described in terms of a set of logical rules describing the diverse levels of the students' performance. The software is simple to use besides being reasonably accurate. Moreover the user friendly interface used in this project turns out to be easy to handle and avoid complications

XI. FUTURE WORK

Possible future work to complete benefit of this work in an educational business institution could be enhancing system functionality by visualizing the system to some available decision support systems, and with good visualization facilities to make results meaningful to educators.

REFERENCES

[1] Han, Jiawei and Micheline Kamber. 2001. Data Mining: Concepts and Techniques. Morgan Kaufman Publishers.

[2] Han,J. and Kamber, M., "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor. 2006.

[3] Alex Berson and Stephen J. Smith, "Data Warehousing, Data Mining, And OLAP", MC Graow–Hill, 1997.

[4] Christopher J. Matheus, Gregory Piatetshy–Shapiro and Dwight Mcneill", Selecting and Reporting what is Interesting The Kefir Application to Health Care Data", Advances in Knowledge Discovery and Data Mining, AAA1 Press/The MIT Press, 1996.

[5] David Cheung, Vincent T., Ada W. Fu and Yongjian Fv, "Efficient Mining of Association Rules in Distributed Databases", IEEE, 1996.Graig Silverstein, Sergey Brin and Rajeev Montwani,

[6] "Beyond Market Baskets: Generalizing Association Rules to Dependence Rules", Data Mining and Knowledge Discovery, Vol. 2, No. 1, Jan 1998, Kluwer Academic Publishers.

[7] Jiawei Han, Laks V. S. Lakshmanan and Raymond T.NG, "Constraint-Based Multidimensional Data Mining", IEEE, August 1999.

[8] Ming-Syan chen, Jiawei Han and Philip S. Yu, "Data Mining: An Overview From a Database Perspective", IEEE Transactions on Knowledge and Data Engineering Vol. 8, No. 6, Dec. 1996.

[9] Rakesh A. grawal, "Parallel Mining of Associations Rule", IEEE, Dec 1996.

[10] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "The KDD Process For Extracting Useful Knowledge From Volume F Data", Communication of ACM, Nov 1996 / Vol. 39, No. 11.

[11] Ussama M. Fayyad, Gregory Piatetsky – Shapiro and Padhraic Smylh, "From Data Mining to Knowledge Discovery, an Overview", Advanced in Knowledge Discovery & Data Mining, AAA1 Press / The MIT Press, Massachusetts Inst. of Tech, 1996.

[12] Y. Balaji Padmanabham and Alexander Tuzhili, "A Belief Driven Method for Discovering Unexpected Patterns", the 4th International Conference Knowledge Discovery and Data Mining, August 1998.

[13] Y. Gauten Das, King-Ip Lin, Heikki Mannila, Gopal Renganathen and Padhrik Smyth, "Rule Discovery From Time Series", Proceedings of The 4th International Conference on Knowledge Discovery and Data Mining, November 1998.

[14] Feelders. A.. H. Daniels and M. Holsheimer. 2000. Methodological and practical aspects of data mining. Inform. Manage.• pp: 271-281.

[15] Agrawal, R., Imielinski, T., and Swami, A., 1993. Mining association rules between sets of items in large relational databases. *Proceedings of ACM SIGMOD international conference on management of data* 1993, 207-216.

[16] R. Agrawal and R. Srikant. Mining sequential patterns. In P. S. Yu and A. L. P. Chen, editors, Proc. 11th Int. Conf. Data Engineering, ICDE, pages 3–14. IEEE Press, 6–10 1995.