Discovery of students' academic patterns using data mining techniques

Mr. Shreenath Acharya
Information Science & Engg department
St. Joseph Engineering College
Mangalore, India
shree.katapady@gmail.com

Ms. Madhu N
Information Science & Engg department
St. Joseph Engineering College
Mangalore, India
nmadhu.nayak@gmail.com

Abstract - **Knowledge discovery is an emerging field which combines the techniques from mathematics, statistics, algorithms and Artificial Intelligence to extract the knowledge. Data mining is a main phase of Knowledge Discovery in Databases (KDD) for extracting the knowledge based on the patterns and their correlation by the application of appropriate association rules to the informations available from the data set. The outcome of the KDD is used to analyse or predict on the future aspects in any area of considerations. In this paper we propose an analysis and prediction of students placements based on the historical informations from the database by considering the students information at different confident levels and support counts to generate the association rules. The widely used algorithm in data mining ie, apriori algorithm is specifically considered for the extraction of the knowledge.**

Keywords: KDD, Data mining, analysis, rules

1. Introduction

The explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge [6]. Today, techniques for the discovery of patterns hidden in large data sets, focusing on issues relating to their feasibility, usefulness, effectiveness and scalability are present. Hence, the proposed system aims at discovery of knowledge from a huge database. Patterns of interest will be mined which will prove useful in near future using available techniques.

The proposed model is intended to mine the interesting patterns from the student academic records and placement database. Nowadays, organizations as well as students need to take decisions appropriately to satisfy their academic requirements. An educational organization should be aware of the trends of placements or recruitments in their institution. Even the students, who would like to pursue any course in particular institution, would want to know the placement trends of that institution. Therefore, the useful patterns that are generated would guide the academicians and the students to make decisions more confidently. Moreover, the patterns generated from student academic records will help the institution to analyse the academic performance of the students. Hence, this work would behave as a decision support tool for the management system to improve their policy-making, setting new strategies and having more advanced decision making procedures.

The remaining section of the paper is organized as follows. Section II explores the various development and their reviews, section III describes the model architecture of the proposed system, section IV describes the methodology & implementation, section V provides an analysis of the obtained results and section VI draws the conclusion and the future scope.

2. Literature Review

The developments in the computing along with their specific requirements has led to increased requirements of large volumes of the complex data. It will be cumbersome to analyse those datas manually as per their need in

an application. Thus data mining plays a vital role in extracting the hidden information from the large data sets (databases).

Faouzi Mhamdi et al [1] described the Knowledge discovery saying that it provides new concepts or concept relationships hidden in large volumes of raw data which has the capacity to automate complex search and data analysis tasks. They also specified that data mining extracts nuggets of knowledge to be used in verification of hypothesis or the prediction and knowledge explanation and its different phases are data preprocessing, data processing or data mining and data post processing.

Usama Fayyad [2] presents an overview of KDD and Data mining in the applications depicting that they are at the intersection of several disciplines including statistics, databases, pattern recognition/AI, visualization, high performance and parallel computing. He also describes the significance of Data Mining which provides techniques that allow managers (higher officials) to identify the valid, useful, understandable correlation and patterns of data.

Mykhaylo Lobur et al [3] provided an insight into the different repositories of information like data warehouses, transactional databases; relational databases etc. leading to varied approaches for the data mining. They also discussed about the four basic mining approaches supported by different mining technologies: predictive model creation supported by supervised induction techniques, link analysis by association & sequence discovery techniques, DB segmentation supported by clustering techniques, and deviation detection supported by statistical techniques. They described Knowledge Discovery as a process to extract the specific patterns of interest from the data using varied data mining algorithms.

Timo Horeis et al [4] described the capability of the conventional systems for knowledge discovery and data mining revealing their ability to extract valid rules from huge data sets. These extracted rules describe the dependencies between attributes and classes in a quantitative way. They also discussed the effect of fusion of this knowledge combined with the qualitative knowledge from several experts' resulting in more comprehensive knowledge about an application area.

M. Lobur et al [5] defined KDD as a process of discovering useful knowledge from the data and data mining as a step in it. They gave an insight into the main goal of KDD as extracting high-level knowledge from low-level data in the large data set context. They also revealed the structure of KDD software systems often embedding statistical procedures for sampling and modeling data, evaluating hypothesis and handling noise within an overall knowledge discovery framework.

Qian Wan et al [7] discussed about the practical usefulness of data mining techniques necessitating an approach which would hope to bring a promising avenue to look at the data from a new angle in order to allow us find new, useful and actionable patterns.

Li zhu et al [8] described the association rules in the context of data mining concentrating only on improving the efficiency of the algorithm neglecting the users understanding and participation. They revealed the fact that students historical records stored in the university databases could be a data source for mining the students' subjective interest and interest degree so that their performance may be improved through teaching and personal trainings.

The proposed data mining model would eventually help educational institutions to improve their decision-making approach.

3. System Architecture

Institutions do not rely on any system to predict the performance of their students' placement. This may result in students becoming unproductive professionals as they are unable to identify their field of interest. So it would be better if we have a system to help educational institutions to produce qualitative students by providing them appropriate guidance. The discovery of the student behaviour patterns assists not only the institution but also the students to help them take right decisions. To justify, if we consider the present educational institutions, we do not find any system which finds patterns of various kinds on the performance of the previous students. Manually doing such a task of finding patterns would be erroneous as large amount of data has to be surfed in order to come to a conclusion. Making the system think would help to take decisions at a faster pace. Thus, the knowledge discovery process with its different phases will prove to be an effective and efficient mechanism to extract the patterns of interest from a database which could lead to right decisions by the concerned persons on any area of their considerations.

Figure 1 describes the process of knowledge discovery as an iterative sequence of the following steps:

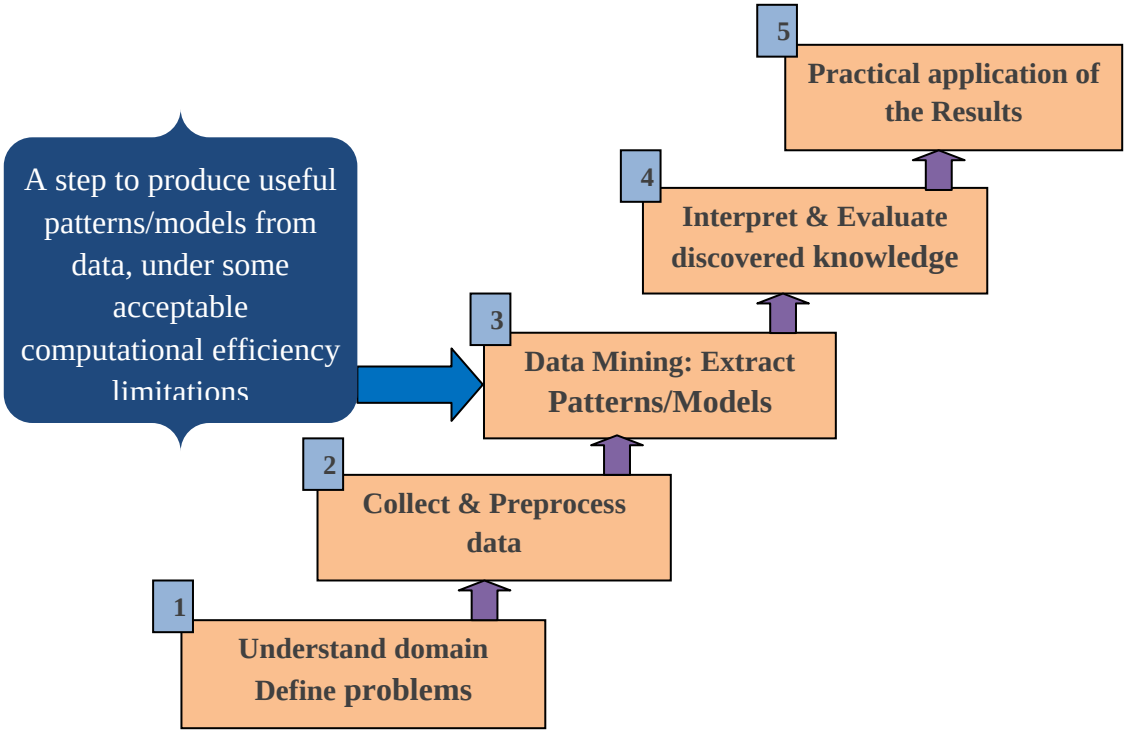

Figure 1. The Knowledge discovery process

1. **Understanding the domain and problem definition:** This is the phase where the problems will be identified and defined to be applied to the KDD process.

2. Collection and preprocessing of the data: This phase collects the informations from the database or dataset and converts it into a standard format to be able to perform the processing.

3. Data Mining: Extraction of the patterns/models: This phase does the main function of KDD ie, it extracts the specific patterns of interest by the application of suitable algorithms (like apriori here).

4. Interpretation & evaluation of the discovered knowledge: This phase uses a domain expert to analyse and interpret the generated patterns.

5. Putting the results in Practical applications: This phase puts forth the knowledge from the previous phase into practical use in applications.

Our system aims at mining attractive, unnoticed, and useful patterns from a database which has been gathered from the placement section of a particular institution. Once done with the identification of patterns, it can be applied to future set of data in order to provide appropriate guidance and counselling to the students. Similar idea can be applied to any other databases to arrive at conclusions that would be beneficial to the institution or organisation. The main intention is to make the system think.

The users of the system are the executives who are having the authority to take decisions on behalf of the students. Even the students can take necessary guidance from the system with the help of their in-charge. For e.g.: The HOD of the departments can make use of the mined patterns to send students to companies suiting their status.

Among the different phases considered for the Knowledge discovery in databases, data mining is the main phase which considers the preprocessed data as the data source. The data source is mined using the apriori algorithm to extract the specific pattern of interest so that it would be analysed by a domain expert to provide suggestions towards the decision making.

A typical data mining system contains the following components [9].

1. Database, Data warehouse: There can be one or a set of databases, data warehouses. Data cleaning, data integration and filtration techniques may be performed on the data.

2. Database or data warehouse server: It is responsible for fetching relevant data based on the user's data mining request.

3. Knowledge base: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included**.**

4. Data mining engine: This is an essential part of the data mining system which consists of functional modules for tasks. Here, it is association analysis.

5. Pattern evaluation module: This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns.

6. User interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task. In addition, this component allows the user to browse database and also query it.

4. Methodology

The different tasks of the data mining has been divided into pre-processing, processing and post-processing.

4.1 Pre-processing

The data will be pre-processed before applying the apriori algorithm to mine patterns.
The pre-processing stage includes: Filtration/Cleaning, Homogenization, Generalization and Integration.

4.1.1 Filteration

The database collected has 21 tables. The drop table query has been applied to filter the database contents so that it only contains the tables 5 tables named comp_info, course, enrolled_in, is_placed, stu_info.

Each table is analysed individually to remove unwanted attributes by the application of specific queries to the contents of the table so that it suits our requirement in the proposed application.

4.1.2 Homogenization

The data in the table has to be brought to a standard form to make the mining process easy.
This process is called as homogenization or the modularization of data.

The 5 tables taken into consideration have been brought into a standard form by eliminating the spelling mistakes, different notations for the same place etc. by applying specific queries so that the mining process becomes easier.

4.1.3 Generalization

The queries have been executed appropriately at both UG and PG level to generalize it for the specific range of salaries.

4.1.4 Integration

All the tables that have been filtered, standardized and generalized are merged by the application of specific queries so that it results in a single table to be fed as the data source for the algorithm.
This merged table is converted into a market basket type of data.

4.2 Processing

The processing stage involves the application of the apriori algorithm for the given support and confidence in order to extract the frequent item set thereby generating the rules of our interest to be analysed for future applications.

Support determines how often a rule is applicable to a given data set. A low support rule is likely to be uninteresting from a business perspective. So, support is often used to eliminate uninteresting rules.

$$\text{Support}(x) = \frac{\text{Number of transactions containing x}}{\text{Total number of transactions}} \quad (1)$$

Confidence determines how frequently items in Y appear in transactions that contain X, when the association is of the form X → Y. It measures the reliability of the inference made by a rule. For a given rule X → Y, the higher the confidence, the more likely it is for Y to be present in transactions that contain X. Confidence also provides an estimate of the conditional probability of Y, given X.

$$\text{Confidence}(x \rightarrow y) = \frac{\text{Number of transactions containing x and y}}{\text{Number of transactions containing x}} \quad (2)$$

4.2.1 Algorithm to calculate FrequentItemsets:

```
set count = candidates size
  read transaction file to a FileInputStream and put it in a BufferedReader
       for(i=0 to number of transactions)
              read  i^th item from transaction file, tokenize it with respect to itemseparator
               for(j=0 to number of items)
                        set Boolean array = true after comparison with oneval[j] where transaction
                        is  present and false otherwise
          for(c=0 to Candidates size)
                    set match = false
                       tokenize candidates (c), store in tokenizer st
                           while(st has more tokens)
                                    store the value in trans[next token of st -1]
                                 if(match)
                                          increment count[c]
          for(i=0 to Candidate size)
                    if((count[i]/number of transactions ) >= minimum support)
                              add candidates at i to frequentCandidates
        reinitialize Candidates with frequentCandidates
 clear frequentCandidates
```

4.2.2 Algorithm to generate rules

```
for all the frequentItemsets formed
          if( frequentItemset !=1)
                    find the combinations formed from that frequent itemsets
                       for each combination
                          generate antecedent, consequent of the rule
                                    confidence = support(antecedent and consequent)/support(confidence)
                     display rules with (confidence > minimum confidence)
```

4.3 Post-Processing

In this stage of post processing, the domain expert analyzes the rules according to the interestingness of the rule. The knowledge base of the user is applied here to filter out the rules which are not interesting.

Association analysis algorithms have the potential to generate a large number of patterns. As the size and dimensionality of real, commercial databases can be very large, we could easily end up with thousands or even millions of patterns, many of which might not be interesting. Sifting through the patterns to identify the most interesting ones is not a trivial task because "One person's trash might be another person's treasure". It is therefore important to establish a set of well-accepted criteria for evaluating the quality of association patterns.

The first set of criteria can be established through statistical arguments. Patterns that involve a set of mutually independent items or cover very few transactions are considered uninteresting because they may capture spurious relationships in the data. Such patterns can be eliminated by applying an objective interestingness measure that uses statistics derived from data to determine whether a pattern is interesting.

The second set of criteria can be established through subjective arguments. A pattern is considered subjectively uninteresting unless it reveals unexpected information about the data or provides useful knowledge that can lead to profitable actions. For example, the rule {BTECH}→{Male} may not be interesting despite having high support and confidence values, because the relationship represented by the rule may seem rather obvious. On the other hand, the rule {cgpa_6}→{comp_add_UP} is interesting because the relationship is quite unexpected and may suggest a new trend in placement. Incorporating subjective knowledge into pattern evaluation is a difficult task because it requires a considerable amount of prior information from the domain experts.

Pattern evaluation is done by incorporating subjective knowledge. The approach used is visualization where it requires a user-friendly environment to keep the human user in the loop. It also allows the domain expert to interact with the data mining system by interpreting and verifying the discovered patterns.

5. Implementation

The implementation of the algorithm and the user interface for the human interaction has been carried out using Java programming language and Netbeans IDE 6.5 is used as an Integrated Development Kit. The NetBeans facilitates the execution of the application in any platform like Windows, Mac OS, Linux and Solaris. We have considered Windows platform and Microsoft SQL Server 2005 as the back end database. The XAMPP which is an open source cross-platform web server is also considered to support the capability of serving dynamic web pages.

6. Results and Analysis

The pre-processing of the database is done manually by applying our domain knowledge with the help of SQL queries. The results obtained by the queries were tested during the process itself by referring back to the original database. So, separate test cases are not used to test the task of pre-processing.

The Apriori Algorithm used in the system to discover student academic behaviour is a well tested algorithm in itself. But, testing of individual phases of the algorithm during the frequent item set generation and rule generation module have been successfully conducted so that it satisfies all the requirements for the analysis in pattern evaluation to be presented to the user.

6.1 Analysis of the rules

6.1.1 Discovered Student Behaviour from the Placement Database

Figure 2 and Figure 3 describe the generation of frequent item set and rule for the sample input shown with support count =15 and the confidence = 8.

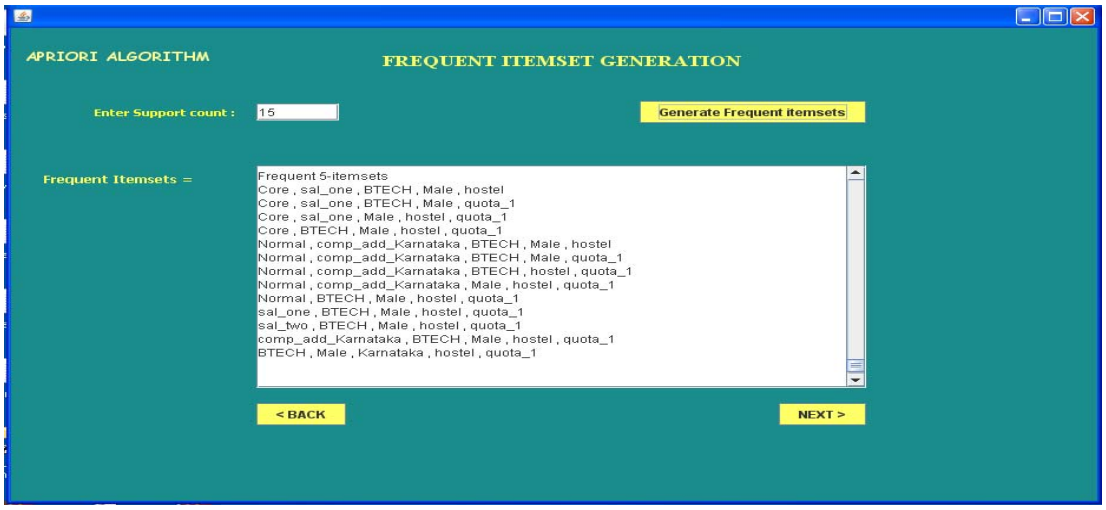

Figure 2. Frequent Itemset Generation

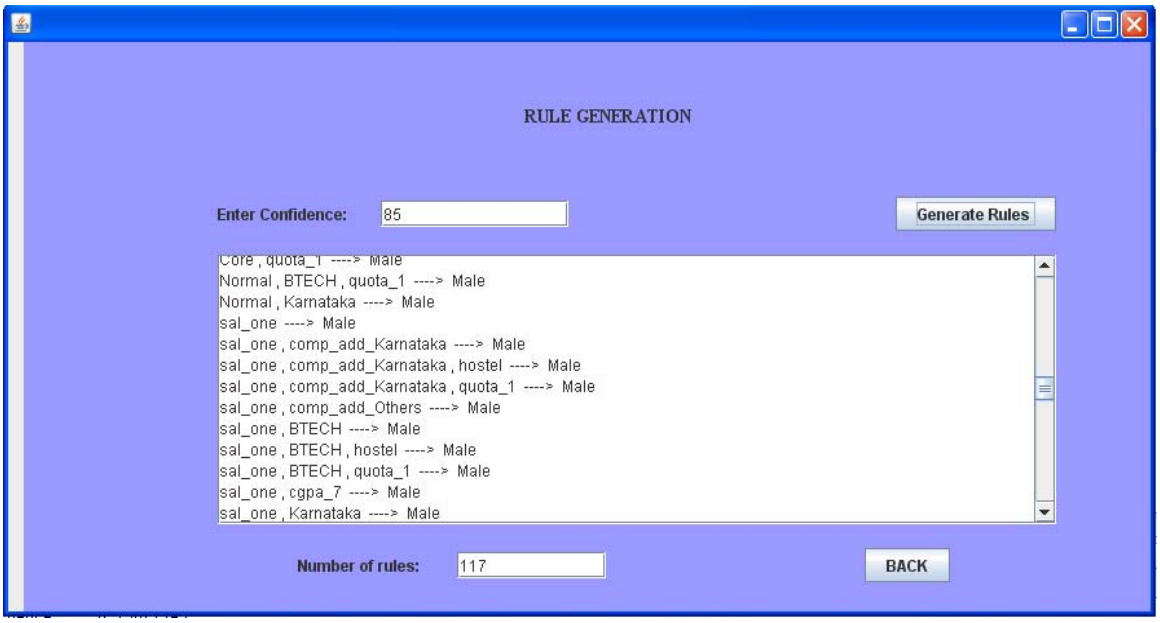

Figure 3. Rule Generation

It is observed that the rules generated when support is low and confidence is high are more interesting.

The support values below 10 and above 40 are giving spurious rules.

The confidence values below 30 and above 90 are giving rules which are not very interesting.

So the favourable range for support is between 10 and 20 and that of confidence is from 70 to 90.

Some of the generalizations considered before analysing the results are:

quota_1 : The students coming from general merit

quota_2 : The students belonging to certain minorities

quota_3 : SC/ST students

quota_4: Physically challenged students

sal_one: The salary range between 0 to 3 lakhs

sal_two: The salary range between 3 to 6 lakhs

sal_three: The salary range between 6 to 11 lakhs

cgpa_6: The cumulative grade point average is greater than or equal to 6 and less than 7

cgpa_7: The cumulative grade point average is greater than or equal to 7 and less than 8

comp_address_others: Faridabad, Goa, Gujarat, Jamshedpur, Kolkata, Rajasthan, Jaipur

The rules that are obtained from the apriori algorithm are analysed as follows for various support and confidence values.

The rules having single element in the antecedent and consequent are pruned.

The test results obtained for some values of support and confidence are as in the table 1 below.

Table 1. Sample test results for different support and confidence values

Sl. No	Support	Confidence	Number of rules generated
1	15	80	199
2	15	85	117
3	15	60	588
4	5	90	872

Similarly the test results are obtained for different support and confidence values and the rules that are unrealistic and uninteresting are eliminated.

A sample of the suggestions to the institution from the obtained patterns:

1. Send male students doing BTECH, and who want to get placed in places like Faridabad, Goa, Gujarat, Jamshedpur, Kolkata, Rajasthan and Jaipur to core companies for placement.

2. Send students doing BTECH, from general merit quota, staying in the hostel and want to get placed in Karnataka to core companies for placement.

3. The male students from Karnataka doing BTECH get placed in a company in Karnataka.

4. Send male students doing BTECH who are from the general merit quota, staying in the hostel and want to get placed in a company in Delhi, to core companies for placement.

5. Male Civil students staying in hostel should be sent to core companies.

6. Students expecting a job in UP with a salary in the range 0 to 3 lakhs, should try for normal companies.

7. Male students doing BTECH, from the general merit quota and staying in the hostel, from the CS branch and like to get in a company in Karnataka, should try for normal companies.

8. MTECH students who are male and hostelites or from quota_1 should be sent to companies paying within 3 lakhs.

Similarly, many more conclusions can be drawn and actions taken for different combinations of support and confidence.

7. Conclusions and Future Scope

Successful Knowledge discovery and Data mining applications play a vital role in extracting unknown knowledge from vast data sets. In this paper we have developed a model for the placement database so that institutions can use it to discover some interesting patterns that could be analyzed to plan their future activities. It has been found to be very useful to the higher authorities like principal, head of the department or the

placement officer for taking decisions to sort out the students based on their educational stream, location, interest for proper management and training leading to their successful career. The issues regarding the outcome of the research is that, it depends on the completeness and the accuracy of the data being analyzed and it requires a domain expert to evaluate the generated rules.

The future scope could be dynamic updation of the database by the user before applying the algorithm to generate the patterns of interest rather than performing it on statistical historical data. Moreover the data mining can be made as constraint-based mining wherein the rules can be generated based on the constraints provided by the user.

References

[1] Faouzi Mhamdi, Mourad Elloumi, "A New Survey On knowledge Discovery And Data Mining" December 2007.

[2] Usama Fayyad, "Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases" Proceedings of the Ninth International Conference on Scientific and Statistical Database Management IEEE Computer Society Washington, DC, USA, 1997.

[3] Mykhaylo Lobur, Yuri Stekh, Vitalij Artsibasov, "Challenges in knowledge discovery and data mining in data" MEMSTECH'2011, 11-14 May 2011, Polyana-Svalyava (Zakarpattya), UKRAINE.

[4] Timo Horeis, Bernhard Sick, "Collaborative Knowledge Discovery & Data Mining: From Knowledge to Experience" Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007).

[5] M. Lobur, Yu. Stekh, A. Kernytskyy, Faisal M.E. Sardieh, "Some Trends in Knowledge Discovery and Data Mining" MEMSTECH'2008, May 21-24, 2008, Polyana, UKRAINE.

[6] Han J, Kamber M, "Data Mining: Concepts and Techniques", Simon Fraser University, Morgan Koufmann Publishers, Second Edition, 2007.

[7] Qian Wan, Aijun An, "Transitional Patterns and their significant milestones" 7th IEEE International Conference on Data Mining, pp. 691 – 696, 2007.

[8] Li zhu, Yanli Li, Xiang Li, "Research on Early – Warning Model of students' academic records based on association rules", Proceedings of World Congress on Computer Science and Information Engineering, pp. 121 – 125, 2009.

[9] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to Data Mining", Pearson Education Inc., Fourth Edition, 2009.