International Journal of Research in Advent Technology, Vol.4, No.4, April 2016
E-ISSN: 2321-9637
Available online at www.ijrat.org

Efficient Implementation of Clustering using Segmentation- for Placement Data

Neel Chaudhari[1], Pritesh Burngule[2], Premnath Borkar[3], Dikshant Chavan[4], Prof. Niti Desai[5]

[1, 2, 3, 4, 5] *Department of Computer Engineering*
[1, 2, 3, 4, 5] *MCT Rajiv Gandhi Institute of Technology Mumbai-53*
Email: neelchaudhari6@gmail.com[1], burngule@gmail.com[2], premnath.borkar@gmail.com[3], dikshantchavan@gmail.com[4], nitidesai@gmail.com[5]

Abstract- Every year the college data goes into terabytes and petabytes as well. There are numerous students who take admission in engineering in India. After completion, some go for Higher Studies while the remaining opts for the campus placement. There is a relationship between the students placed according to the past records. The students placed according to their past records, is a least explored area which really needs to be focused. The system allows the user to get an idea for which organization he might be placed. The data from the placement cell as well as the exam department is collaborated to get the association between them. The system finds the association between student's performance and the company he is placed in. It will be a global system which can handle big data efficiently and give approximately correct output for data mining purposes.

Keywords: Placement Data Mining; Clustering; Segmentation; K-means etc…

1. INTRODUCTION

In today's world, data in huge quantities is being accumulated in the data repository. The data constructed usually has a huge gap from the stored data to the knowledge. That's where Data Mining comes into picture this change won't occur automatically. One of the most desired attributes of Data Mining is gathering knowledge from massive data. This vast amount of data is mined by a number of Data Mining techniques such as association, clustering, classification. Data mining focuses on structured data, such as relational and transactional data [1].The objective of data mining is to identify valid novel, potentially useful, and understandable correlations and patterns in existing data [2]. Placement data can include student's percentage, skill sets, company's name, company's package also. Therefore, observation using parameters like student's skill sets, percentage and company's criteria provide the prediction about the pay package.

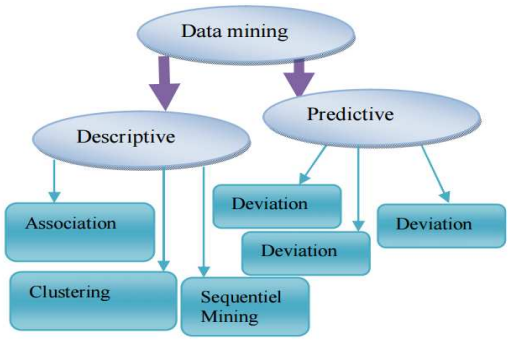
1.1 DATA MINING TECHNIQUES

Data mining which is also known as Knowledge Discovery from Databases (KDD) is defined as the activity of extracting knowledge from huge amount of data and identifying patterns accordingly. Therefore, data mining is divided into three crucial tasks as given below:

Classification: Classification refers to the way of Organizing data in given classes. It classifies the data according to its categories. It is a form of supervised learning. It is a kind of learning wherein a training set of correctly identified observations are available.

Clustering: Clustering is collecting data objects and grouping them as a single type cluster by taking similar objects to one another within the same clusters and dissimilar to the objects related in other clusters. Initially, we have no idea that how many numbers of groups are present. It is a form of unsupervised learning. Clustering uses a process of differentiating by finding similarities between data according to their characteristics.

Association: Association analysis is the exploration of what are commonly termed as association rules. It includes the study of the frequency of items occurring together in transactional databases and based on an identifier called threshold/support which identifies the frequent item sets. To find association between multiple attributes, generate rules from data sets, the data can be used. Therefore, this task is known as association rule mining. In a certain given set of transaction, the prediction of the occurrence of an item can be found using rules based on the occurrences of other items in the transaction. The main task of association rule mining is to find all rules having support ≥ minimum support threshold and confidence ≥ minimum confidence threshold [3].



1.2 ASSOCIATION RULE MINING

This part of data mining focuses on analysis of data for identifying occurrences of events and hence uses the value of support and confidence. It is used for searching relationships among items in given data set [4].

BASIC TERMINOLOGY:

1. Tuples are transactions, attribute-value pairs are items.
2.Association rule: {A1,B1,C1,D1,...} => {E1,F1,G1,...}, where A1,B1,C1,D1,E1,F1,G1,... are items [5].

Support: The support of an item set shows how frequently an item set is appearing in a single transaction in the database.

Formula: $I = P(A \cap B) = (A \cap B)/N$
Where N is total number of items
Range: {0, 1}

Confidence: Confidence can be defined as the ratio of the number of transactions that contain both A and B to the number of transactions that contain only A.

Formula: $I = P(BA) = P(A \cap B)/P(A)$
Range: {0, 1}

2. METHODOLOGY

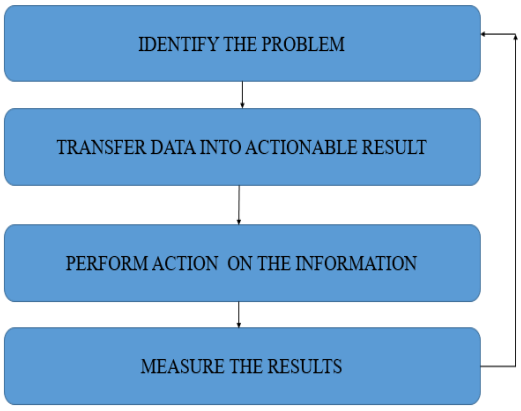


Fig 1: Flow of Data Mining

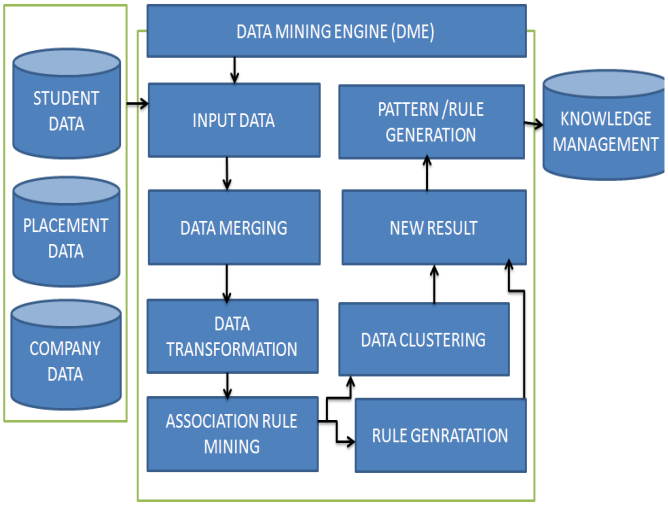


Fig 2: Data Mining Engine

1. Data merging-: Data collected is merged and sorted according to the needs of our project. Data such as result, pay package in taken into consideration for our project. The file created is then converted into arff format.
2. Pattern Generation-: Derive the patterns received from the output generated. Classify the patterns by checking the results inputted of various students. Derive the companies a particular student is eligible for.

3. SEGMENTATION:

Segmentation is a process of dividing a student database in a form of individualism which are same in specific ways applicable to college placements, based on parameters like percentage, branch, interests, extra-curricular skills, and so on. A segmentation technique which is targeting as a group of students with specific criteria [6].

4. FEATURES

Feature of a system is what makes it powerful and attractive system. As no one has ever developed such a system, it is one of its kinds.
Following are the features of the proposed system:

a) 1st placement analyzer:
It's the first placement analyzer ever made in the history of Rajiv Gandhi Institute of Technology and it has some features too.
b) Supports any File Format:
It supports any kind of file format, but should be well formatted so the system can interpret the data properly.
c) Fast :
As we are using clustering and segmentation techniques, it is easy to access the data due to formation of similar clusters and gives much faster output than any other method.
d) Transparency:

The system is transparent to admin and student and the transparency is maintains in such a way that the admin is given more privileges than student to access the system.

e) Distribution:

As the system possess distributed property, the data is distributed among several clusters according to its properties.

f) Flexibility:

Our system is flexible with other data as well and the necessary changes in the future can be made. Hence it can be used in different sectors.

5. PROPOSED SYSTEM

The proposed system is used in Rajiv Gandhi Institute of Technology (Engineering College) it can be used in other engineering and non-engineering colleges too.

The other option is using it in the schools and classes to predict the upcoming grades of a student for improvement and enhancement of the student in getting good grades.

It can be used in other sectors too, such as:

- Retails
- Education
- Medical Education
- Marketing
- Business Strategies
- Market Basket Analysis

Student's marks entry and percentage is in a convenient manner. Database is stored in excel sheets with the number of students which were allowed to sit for the campus placements depending upon the criteria and the companies in which they were placed. This database is highly scalable as N number of student entries can be made without disturbing the performance of the system. It is also good for performing complex queries which may be later used for data mining purposes. As excel sheet is used which is easy to access .Our system is highly reliable as it is good to execute concurrent execution of process and performance is high. A new staff using the software will take no time to get familiar with the system as the system is not very much complex, it is user friendly.

6. FUTURE SCOPE

Every system has its own limitations. Following are some limitations of our system:

The scope defines the boundary of a system. The system cannot work with other software. But it has the ability to accept file in any format but should be well-formatted to interpret. The system is not available in other languages other than English. The data is inputted need to be in proper format or else the processing of the data won't happen and it will abort the process. Also the process take a lot of memory while executing the process as It forms the clusters and segments of the candidate set. The records of the students are stored in excel sheets. The data can be fetched from excel sheets if needed. Forecasting is more easier in excel sheets. By using data mining technique i.e. linear Regression which will help in forecasting students result. We can find

Out some particular features for e.g. predict the companies and their package depending upon student's percentage and extra-curricular skills. Divide the students into clusters according to the percentage and companies by their package. Also for analytical study various graphs, pie-charts can be generated accordingly. Report predictions are one click away. Various reports of students based on the company's requirements can be generated very easily. Suppose a student, who needs a placement information of the company for example list of students having placed in academic year 2015-16 depending on the marks can get it easily and need not search it manually.

7. CONCLUSIONS

In this proposed system, a system is developed which will ease the use of the student's research about the placement in finding their dream jobs and it will also indirectly help students to avail the eligibility and predictions at the earliest. This is done by making the GUI very user-friendly to use for both admin and users and also putting complex conditions in a simple way which would increase the performance of the system, thus automating the system as much as possible. The system also takes into account the list of placed students in that year .Though the redundancy is present, but due to an efficient database being used, it is scalable and performance is not reduced.

ACKNOWLEDGEMENT

We would like to thank our guide, Prof. Niti Desai, Department of Computer Engineering, for all the advice, Encouragement and constant support that she has given throughout our project work. We would also like to thank Prof. D.S.Suralkar and Placement cell for their support and valuable suggestion has made our project achievable.

REFERENCES

[1]. Deepika Raj K, , and Saani H. "Semi-automatic building of Domain Module by use of novel machine learning approach", 2015 International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS), 2015.

[2]. Mathai, Paul P. and Balan, R. V. Siva. "Optimizing The Achieved Frequent Item Sets Using Genetic Algorithm", International Journal of Applied Engineering Research, 2015.

[3]. paper.ijcsns.org

[4].www.ijirae.com

[5].www.iosrjournals.org

[6].zyxo.wordpress.com