International Journal of Computer Theory and Engineering, Vol. 2, No. 5, October, 2010
1793-8201

Course planning of higher education to meet market demand by using data mining techniques – a case of a Technical University in India

C.Usharani and Rm.Chandrasekaran

***Abstract*—Student performance in University courses is of great concern to the higher education managements where several factors may affect the performance. Data mining techniques are used to analyze the course preferences and course completion rates of enrollees in different Institutions in a Technical University. Courses were classified into three broad groups. Records of enrollees from 2007-09 were then analyzed by three data mining algorithms: Decision Tree, Link Analysis, and Decision Forest. Decision tree is used to find enrollee course preferences, Link Analysis is used to determine the correlation between course category and enrollee profession (part time), and Decision Forest is used to find the probability of enrollees completing preferred courses. Results will be used as a reference for future curriculum development.**

***Index Terms*—Data mining, Decision tree algorithm; Link analysis algorithms; Decision forest algorithm**

I. Introduction

In the last two decades, a revolution in education has started. The number of institutions of higher education has increased steadily upgrading to universities / institutes of technology offering 4 year degree programs. As the number of colleges has increased, maintaining the database is big problem. colleges and universities have found themselves facing stiff competition for students. As a result, many colleges and universities have established different mode of courses. The purpose of this research was to use data mining techniques to uncover the preferences and future choices of students of a Technical University. This data would then be used to better target curriculum on student needs. Decision Tree Algorithms, Link Analysis Algorithms, and Decision Forest Algorithms developed from the theory of Data mining were used in this research. Data consisted of the student records of enrollees in courses in the 2 academic years from 2007 to 2009 Results of the research included (1) course preferences of enrollees; (2) the relationship between course categories offered and enrollee profession; (3) the probability of course completion. These results will be used as a reference in future curriculum development at the University.

Anna University Tiruchirappalli, Tiruchirappalli, Tamilnadu, India.
Corresponding author: Tel: 9443433386, Mail to: ushaji_cs@yahoo.com , aurmc@sify.com

II. Literature Review

A. Scholars definitions

Definitions for data mining. Frawley, Paitetsky-Shapiro, and Matheus (1991) declared that data mining is actually a process of discovering of no obvious, unprecedented, and potentially useful information. Curt (1995) defined data mining as a database transformation process, in which the information is transformed from unorganized vocabulary and number to organized data, and later turned into knowledge from which a decision can be made.

Fayyad, Piatesky-Shapiro, and Smyth (1996) stated that data mining is an uncomplicated process of discovering the valid, brand new, potentially useful, and comprehensive patterns from data. Hui and Jha (2000) defined data mining as an analysis of automation and semi-automation for the discovery of meaningful relationships and rules from a large amount of data in a database.

Peacock (1998) declared that data mining can be categorized as narrow and broad. The narrow definition is limited by the methodology of mechanical learning which emphasizes the discovery process and uses Artificial Intelligence such as Neural Networks, Correlation Rule, Decision Tree Algorithms and Genetic Algorithms.

By contrast, the broad definition emphasizes the knowledge discovery in database (KDD), the process of obtaining, transforming, clarifying, analyzing, confirming and enduring the meaning of the data within existing customers or outside of the cooperation, and then results in a backup system of decision making which is continuously being modified and maintained.

Hand, Blunt, Kelly, and Adams(2000) stated that data mining is a process that discovers Interesting and valuable information from a database. Berson, Smith, and Thearling (2001) argued that the appeal of data mining lies in its forecasting competence instead of merely in its ability to trace back.

Hui and Jha (2000) indicated that the data mining process should include the following .

B. Steps

1) Establishing the mining goals: using domain knowledge to select data relevant to the research goal.
2) Selection of data: identifying the characteristics of variables on which mining can be performed.
3) Data pre-processing: removing noisy, erroneous, and

incomplete data.

4) Data transformation: transforming the data into a new format in order to mine additional information.
5) Data warehousing: the process of envisioning, planning, building, using, managing, maintaining, and enhancing databases.
6) Data mining: discovering correlations among variables after performing data mining and finding interesting, meaningful, and valuable knowledge based on the research topic.
7) Evaluating the mining results: elaborating and evaluating the results after knowledge is obtained. To summarize the foregoing definitions, data mining is a process of obtaining knowledge. The key to the process is comprehension of the research application, and then constructing a data set by collecting data relevant to the research field, purifying the data in the targeted database to eliminate erroneous data, supplementing missing data, simplifying and transforming the data, and at last discovering the patterns and among the data and presenting them as useful knowledge. Artificial intelligence has made remarkable progress in recent decades. A diversity of data mining algorithms has been developed for application to different types of data. Brief descriptions of these algorithms follow.

III. Techniques And Algorithms

A) Decision Tree

B) Link Analysis

C) Decision Forest

A. Decision Tree

This algorithm uses the technique of Information Gain. The data is automatically clustered using a preset significance standard based on the theory of classification standards, which solves the problem of multiple clustering. Decision Tree clusters data by using models like CHAID, CART, C4.5, and C5. The result comes out differently as determined by either the Boolean value or the string. Decision Tree is used to search for clustering rules and for further analysis of clustering results.

B. Link Analysis

This algorithm uses Boolean operations based on Graph Theory. It clusters variables into couples, and calculates the significance and correlation between each couple. Significance shows the edge possibility of each couple; and correlation shows the edge degree between two nodes. The edge degree is presented as the support degree. The results of Link Analysis can be represented as a graph of nodes and edges.

C. Decision Forest

This algorithm clusters selected variables, and obtains the clustering rule from the results. This method begins by multiple clustering according to target variables, and proceeds to cluster the forecasted variables. The continuous clustering to the forecasted variables will result in the discovery of rules for clustering the data. In the process of clustering, the technique of Information Gain is used to compare the data. For instance, doctors can use patient descriptions to construct clustering rules and refine an individualized medical system which will facilitate diagnosis of illness.

1) Information Gain Pseudocode

```
infoGain(examples, attribute,
entropyOfSet)
   gain = entropyOfSet
   for value in attributeValues(examples,
attribute):
       sub = subset(examples, attribute,
value)
       gain -=  (number in sub)/(total
number of examples) * entropy(sub)
   return gain
```

2) Entropy Pseudocode

```
entropy(examples)
   '''
   log2(x) = log(x)/log(2)
   '''
   result = 0
   # handle target attributes with
arbitrary labels
   dictionary =
summarizeExamples(examples,
targetAttribute)
   for key in dictionary:
       proportion = dictionary[key]/total
number of examples
       result -= proportion *
log2(proportion)
   return result
```

3) Classification Techniques

```
classify( example: example_type, root:
node_ptr ) returns class_type
  current_pos = root
  WHILE  current_pos is not a leaf
     IF  example satisfies
current_pos^.criteria
     THEN   current_pos :=
current_pos^.yes_branch
     ELSE   current_pos :=
current_pos^.no_branch
  { ends with current_pos on a leaf }
  return current_pos^.class
```

IV. Methodology

The database consisted of data from enrollees in courses during the period 2007 to 2009 at Technical University Records of all PG students were collected to form the database for data mining. After taking into consideration the capabilities of the eight different algorithms depicted in Table 1 in classification, prediction, clustering and description, Decision Tree, Link Analysis and Decision Forest were selected. From the viewpoint of practical experiments with data mining, variable definition and variable characteristics are far more important than building

models, and more time will be invested in those activities. Prior to the step of data mining in this research, three separate variables were selected, reflecting the goals of this research. They are the **course category**, **completion status**, and **enrollee profession**. The characteristics of three variables are described as follows:

A. Factors that determining the Quality of Education

There are different factors that affect the quality of Education in different situations. How the factors determining the quality of education using data mining techniques. It will increase the performance of students and enhance the Quality of Higher Education.

1) Quality Measures:
   - Parent Income: Is parent income is a factor to determine the quality of education?
   - Environment: Is environment is a factor to determine the quality of education?
   - Community: Is community is a factor to determine the quality of education?
   - School Academics: Is school academis is a factor to determine the quality of education of education?

B. Course category:

Courses over the last two years in the Technical University were categorized into 3 groups:

1) Engineering,
2) Business Management
3) Computer Applications.

Courses in the Engineering category include more than 10 branches. Example Computer Science and Engineering, Software Engineering. Courses in the category of Business and Management are Enterprise Resource Planning (ERP), Marketing, and Human Resources. Computer Applications consists of Applications of Computer

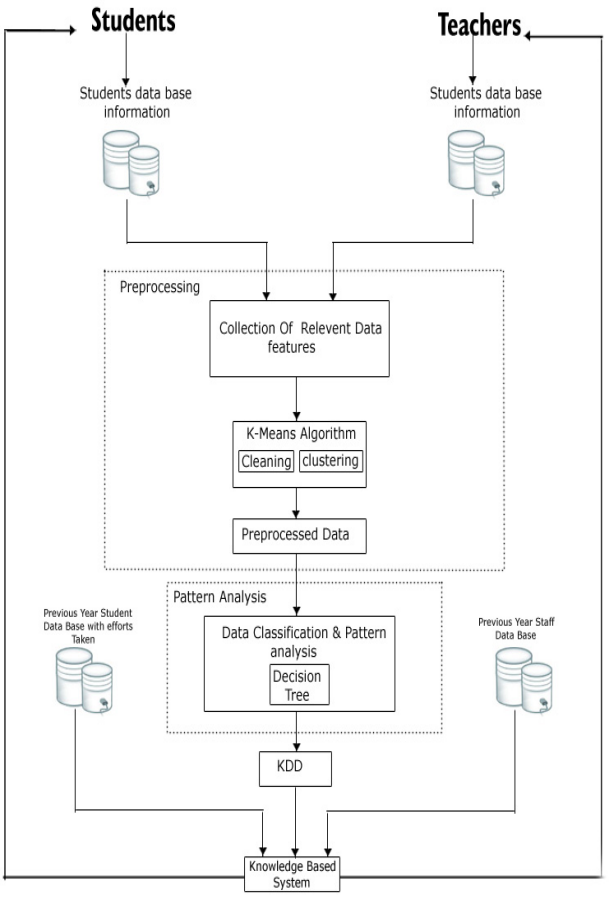
C. Completion status:

It is often the case that planned courses may not be opened because too few students enroll. Additionally, enrollees are often unable to complete the course. Therefore completion status is selected as variable in the research, divided into two groups: enrollees who completed the course, and enrollees who did not complete course.

D. Enrollee Profession:

Technical University in India offers courses in Full-Time students profession is student only. Part Time student profession is and employee of any company. The part time students are having enrollee profession.

SYSTEM ARCHITECTURE



V. CHAID FOR DATA MINING

Decision Tree can automatically split and cluster data based on the preset significance standard, and build up a tree structure from the clustering event. Based on the tree structure, certain rules can be obtained, and the correlation between events found for further forecasting. Each internal node in the tree structure is tested by a preset significance level. Branches are processed by the values of groups or multiple values of groups. This means that the branch of each internal node may have another branch which may also be the internal node of another branch. They are tested in order according to the significance level until the branch cannot be split and comes to an end. The terminal node of the branch is called the leaf node. The path of each leaf node explains the reason for and results of each event. Commonly used models of Decision Tree include Chi-Square Automatic Interaction Detector (CHAID), Classification and Regression Tree (CART), C4.5 and C5. CHAID is used for data mining in this research. Prior to the application of CHAID, the targeted variable and forecasted variable should be defined. In this research, course category, enrollee profession and their characteristic data were taken as the forecasted variables; and enrollment status and its characteristic data were used as the targeted variables to search for preferred courses.

After CHAID was applied, a tree structure was built up based on the targeted variables and forecasted variables (Fig.

1). Completion status is clustered for PG students. (Fig.1). The Root Node represents enrollees of entire PG students completed status. Node one (Node 1), represents completed status of students in Engineering has an 84.39% rate, showing that enrollees are more likely to complete these courses. The other node, (Node 2), represents completed status of students in Management at 56.39%. The other node, (Node 3), represents completed status of students in Applications at 54.93%. Node 1 can be further subdivided into two nodes, Node 4: enrollees in the Engineering sector (completion status of 88.79%), and Node 5: enrollees in the Technology (completion status of 80.79%). With the p-value preset significance level less than 0.05, the third and fourth nodes may each be divided into two internal nodes. The fourth node(Node-4) may be further divided into two nodes, Node 6, and Node 7. These Nodes are called leaf nodes. It reaches at the end. After this, there is no further subdivision. Because it directly reaches the class of specified degree. The Node 6 represents the completed status of students in Software Engineering. (98.35%). Node 7 represents the completed status of students in Computer Science and Engineering (81.88%). Node 5 may be divided into three nodes, Node 8, Node 9 and Node 10. This division reveals that enrollees from Engineering sector However, the completion rate of Node 2 in Business Management is (84.16%) and the completion rate of Node 3 in Computer Applications is (82.57%).The enrollee in Node 2 and Node 3 (Business Management and Computer Applications)

Note: These Nodes may need further sub division, if the two nodes have any section (When the enrollees is high). In Fig. 1, the first (Node 1), the fourth node (Node 4) and the fifth Node (Node 5) are internal nodes of Node 1, and the second (Node 2), the third (Node 3), the sixth (Node 6), the seventh (Node 7), the eight (Node 8), the ninth (Node 9) and the tenth (Node 10) are the leaf nodes. Based on the information derived from the leaf nodes of the tree structure, the correlation between enrollee course preference and enrollee profession is obtained.

VI. Link Analysis For Data Mining

Link Analysis requires setting the degree of closeness. If the correlation value is set to a low value, then all the related data will show a close correlation; whereas if the correlation value is set high, then only highly related data will show a close correlation. Consequently, the setting of the correlation value is the key factor, and its definition should be based upon the best correlation among the data. If the correlation value is set at 6, information regarding the correlation between completion status, course category, and enrollee profession can be found. (1) Enrollees from the Engineering sector, who completed a course, showed a correlation value of 10.3, with a support degree of 533. (2) Enrollees who completed a course in Business Management, showed a correlation value at 8.7, with a support degree of 471. (3) Enrollees who completed a course in Computer Applications had a correlation value of 23.1, with a support degree of 52. The support degree is the number of related data between both characteristics of the variables. For example, among the 1408 records, the total number of enrollees who completed courses and were enrollees from the manufacturing sector is 533, so the support degree is 533. The formula for the correlation value is:

Correlation = -1n (P- value)

where P-value is a probability that the detected correlation is merely a statistical fluctuation. It is used to calculate the degree of significance of each couple according to the Boolean Yes (True) value of each group and every group; and the degree of significance is calculated by 1n based on the proportion of the Cumulative Hyper-geometric Distribution.

VII. DecisionForest For Data Mining

Like Decision Tree, the appointed targeted variables and forecasted variables are defined prior to the application of Decision Forest. In this research, course category, completion status, and the characteristics of those two variables were selected as forecasted variables; enrollee profession and its characteristics were selected as the targeted variable. The purpose of this algorithm is to find the course preferences and course completion rate for the completion status of enrollees from different professions. In Decision Forest, the Engineering, Business Management and Computer Applications sectors are clustered based on the characteristics of enrollee profession. For each sector, completion status is further clustered depending on whether the course was completed.

Completion status is also clustered by course category. The clustering to forecasted variables continues until the rule is found. The split of the details of clustering to the forecasted variable is determined by whether the split criteria value is set high or low. In this research, the split criteria value is set to 4 and 5. When the value is set to 4, the clustering includes the area enclosed by the full line and the dotted line. When the value is set to 5, the clustering only covers the area enclosed by the full line. The results from the application of Decision Forest are: (1) Of enrollees who enrolled in courses of Business and Management completed courses, 42.5% were from the Engineering sector, and 57.5% were from the Computer Application sector completed courses, 51.0%.

VIII. Conclusion

Discussion of the three algorithms In our study three algorithms, Decision Tree, Link Analysis, and Decision Forest, were used to analyze the data from enrollees in Technical University. Decision Tree demonstrated that courses in Engineering, Business Management and Computer Applications are the three most popular courses. Enrollees from the Engineering sector largely take Engineering and Technology courses. By contrast, Business Management courses are most popular with enrollees from the Computer Applications sector.

Link Analysis then showed that there is a high correlation between Engineering and Technology sector enrollees. The

next highest correlation was found between course completion and courses of Business management. At the same time, Computer Applications courses show a high correlation with course incompletion. Finally, Decision Forest shows that enrollees from the Engineering sector primarily take courses in Business Management and Engineering, Application of three algorithms for data mining in this research allows for triangulation of results, which are similar across all three algorithms. Equipped with reliable analytical results, curriculum decision makers will be able to make adjustments in the curriculum mix to better serve student needs. This also indicates the precision of the variables, the usefulness of the variable definitions and their characteristics, and the suitability of the analytical methodology.

In this research, concluded that the paper will enhance the performance of students to improve the quality of higher education by using data mining techniques.

ACKNOWLEDGMENT

Authors gratefully acknowledge the authorities of Anna University Tiruchirappalli for the facilities offered and encouragement to carry out this work. This part of work is supported by second author and my guide.

REFERENCES

[1] C.ROMERO,S.VENTURA, "Educational Data Mining – A Survey from 1995 to 2005", Expert Systems with Applications, Elsevier,33(2007) 135 – 146.
[2] Berry, M. J. A., & Linoff, G. (1997). Data mining technique for marketing sale, and customer, support. Wiley Computer and Sons, Inc.
[3] Berson, A., Smith, S., & Thearling, K. (2001). Building data mining application for CRM. New York: McGraw-Hill Inc.
[4] Curt, H. (1995). The Devile's in the detail: techniques, tool, and applications for data mining and knowledge discovery – Part 1. Intelligent Software Strategies,6(9), 1 – 15.

C.Usharani received the B.Tech. Information Technology from Anna University Chennai and pursuing the M.E. in Software Engineering from Anna University Tiruchirappalli, Tamilnadu. India.

Dr. R. M. Chandrasekaran received the B.E Degree in Electrical and Electronics ngineering from Maduari Kamaraj University in 1982 and the MBA (Systems) in 1995 from Annamalai University, M.E in Computer Science and Engineering from Anna University and PhD Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 1995, 1998 and 2006 respectively. He is currently working as a Registrar in Anna University Tiruchirappalli, Tiruchirappalli Tamil Nadu, India. From 1999 to 2001 he worked as a software consultant in Etiam, Inc, California, USA. He has conducted Workshops and Conferences in the Areas of Multimedia, Business Intelligence and Analysis of algorithms, Data Mining. He has presented and published more than 32 papers in conferences and journals and is the author of the book Numerical Methods with C++ Program (PHI, 2005). His Research interests include Data Mining, Algorithms, Networks, Software Engineering, Network Security, Text Mining. He is Life member of the Computer Society of India, Indian Society for Technical Education, Institute of Engineers, Indian Science Congress Association. World Academy of Science, Engineering and Technology 27 2007.

TABLE 1. CHARACTERISTICS OF SELECTED DATA MINING ALGORITHMS

Algorithm	Classification	Estimation	Prediction	Affinity grouping	Clustering	Description
MBA			X	X	X	X
MBR	X		X	X	X	
Automatic Cluster Detection				X		
Link Analysis	X		X		X	X
Decision Tree	X		X		X	X
Neural Networks	X	X	X		X	
Genetic Algorithms	X		X			
Decision Forest	X		X		X	X

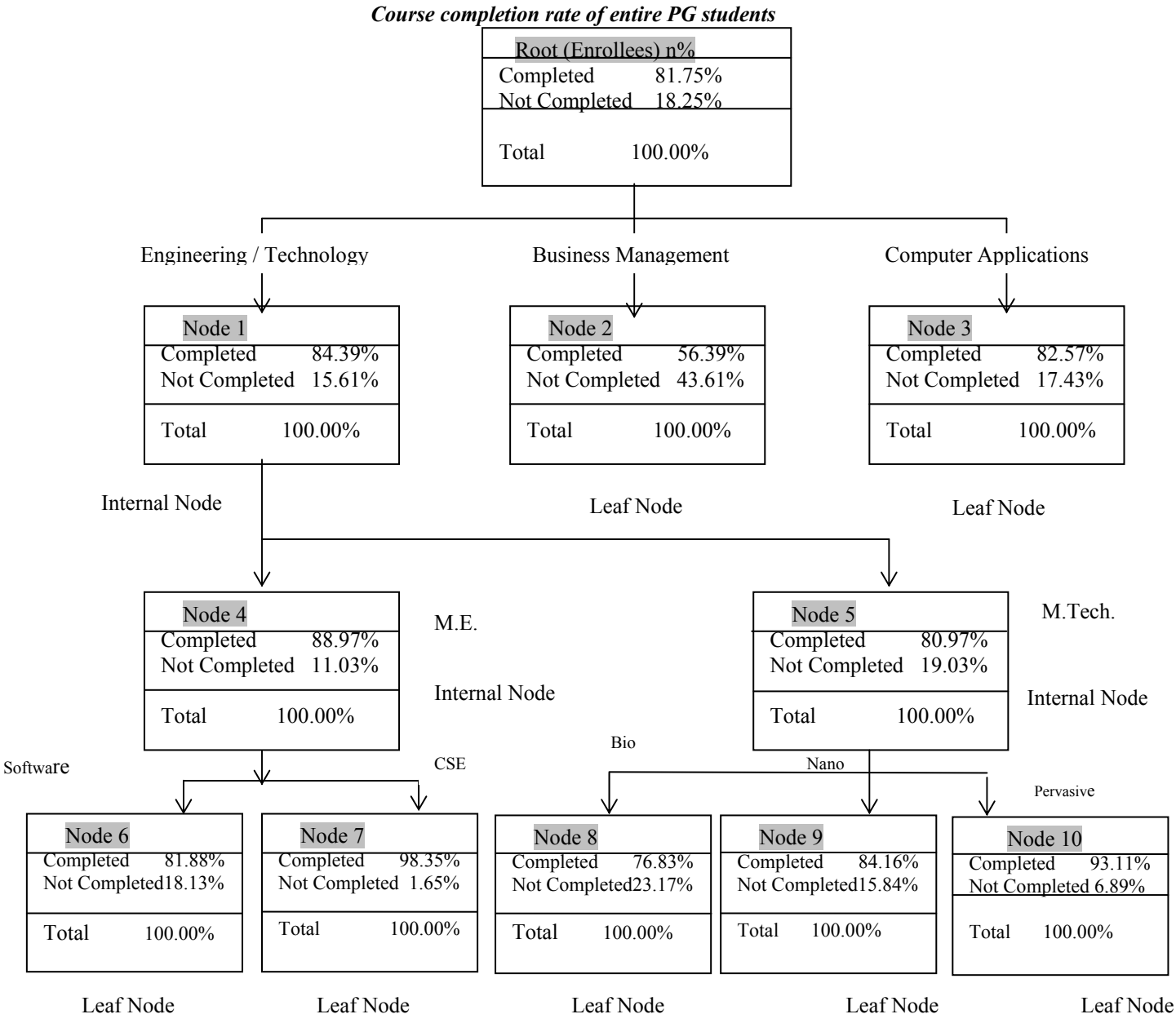


Fig. 1. Structure tree relation of completion status.