College students' Network behavior Using data mining and feature analysis

Thatiparthi Sravani1, Srinivasa Rao Madala2, Sk HeenaKauser3
1 M.Tech., Scholar, Department of Computer Science and Engineering, 
2 Professor & HOD, Department of Computer Science and Engineering, 
3 Asst. Professor, Department of Computer Science and Engineering 
PACE Institute of Technology and Sciences

Abstract. Teachers may use advanced analytics to rapidly and correctly understand undergraduate behavior trends, especially when it comes to identifying undergraduate groupings that need to be focused on at a later time. This study uses data mining cluster analysis to analyze the constituent behavior of 3,245 undergraduates in a specific level 'B' institution's college network. According to the data, there are four different undergraduate groups with different Web access features, with 350 participants using the accomplishments and other variables of their success have an influence on these students. As a result of this research, we were able to collect data on undergraduate college network activity, which may be used to aid in the development of academic advising management.

Keywords—Data mining, Undergraduate Network Behavior; Big Data

1. Introduction

Graduate students' counselors are under-resourced, which results in a reduction of opportunity for primary prevention in many probable difficulties with the administration of the courses, which is a cause for worry. It has been possible for undergraduate advisers to conduct statistical assessments of their behavior from the commencement of this millennium, thanks to the rapid growth of modern digital innovations in academia and the building of digital campuses. Due to the fact that they grew up in an Internet-based world, the Web has a profound influence on their real activities, learning. This makes it possible to study the university network's activity using vast amounts of information. As data types and sizes continue to grow exponentially, educational counselors must find ability to separate meaningful information from them. A mix of Big Data and data mining approaches was used to analyze the network activity of 'B' grade undergraduate college students, as well as the characteristics of the rules controlling connectivity for undergrads. This work might potentially be used to teach students how to mine data [1].
2. Proposed System

2.1 Data capturing and cleansing

Results from this investigation focuses on the utilization of institutional networks and undergraduate behaviors. Basically, there are six elements in the statistics utilized in the college's system: identity, live day, disconnected day, active duration, disconnected duration, inbounds. Each information item captures the sign in information just once, and the total number of data elements for university network use is 2384. There are also a variety of other educational indicators such as exam scores, university card usage, and information on textbook lending. Most of the time, the data that is immediately acquired is imprecise. We cleaned the data before using the data mining technique. In the process of information cleansing, erroneous facts and sensitive records are removed, while pertinent info is retained for statistical references. In this study, the focus is on the features of undergraduate groups rather than the security of undergraduate Web object's confidentiality.

2.2 Indicator network design

Three thousand and forty-five undergraduates were enrolled in this investigation for four years, and only six variables were collected for the initial period: undergraduate identification, digital day, standalone schedule, online and offline times. A quantitative analytical technique is used to clearly and logically describe the characteristics of undergraduate network activity, based on the beginning domains, in this investigation. According to the present education system and educational timetable, each instructional term is separated into four months. There is a problem with reliability performance because to graduating period occurring during last six months of the year. For this reason, 39 half months was studied in this investigation. 28 recurring analytical markers have been used to measure the position and stream of each participant throughout the course of a trimester period (1-25 trimester months). The particular parameters are shown in Table 1.
2.3 Data Preprocessing

When the university campus network was utilized to gather info for this study, it was characterized by a large amount of information, a high intensity of data gathering (info are recorded every time you sign in), and continuous. Operational (slope) information is similar to the information [2]. Let's say you want to interpret information from the university campus network using operational data processing. As a result of the beginning statistical information, the investigation indicator information for each undergraduate during 1-25 trimester weeks of the university campus network should be transformed into operational information.

Table I. Indicator of Behavioral Analysis on University

| Index number | Index Name          |
|--------------|---------------------|
| 1            | Time                |
| 2            | Flow                |
| 3            | AveTime             |
| 4            | AveFlow             |
| 5            | DayTime             |
| 6            | NightTime           |
| 7            | AveDayTime          |
| 8            | AveNightTime        |
| 9            | WeekdayTime         |
| 10           | WeekendTime         |
| 11           | AveWeekdayTime      |
| 12           | AveWeekendTime      |
| 13           | DayFlow             |
| 14           | NightFlow           |
| 15           | AveDayFlow          |
| 16           | AveNightFlow        |
| 17           | WeekdayFlow         |
| 18           | WeekendFlow         |
| 19           | AveWeekdayFlow      |
| 20           | AveWeekendFlow      |
| 21           | Flow/Time           |
| 22           | DayFlow/DayTime     |
| 23           | NightFlow/NightTime |
| 24           | WeekdayFlow/WeekdayTime |
| 25           | WeekendFlow/WeekendTime |
| 26           | Times               |
| 27           | Flow/Times          |
| 28           | Time/Times          |
According to the computation concept, raw data should be processed in a way that preserves content that is readily accessible, and flattening should be used to meet essential The concept of computing Gaussian extrapolation, the flattening core procedure, and the estimation of spline methods are all appropriate practices for pre-treatment [3]. To evaluate 22 sets of data points from 2,512 students on the investigation indicator, the investigation used a spline method. The R-studio application was used to create the R-language programming source code. Using a B-spine basis functionality, flat curvature information was estimated, smoothing data was converted into operational data, and operational data was then subjected to principal component analysis. On the premise of the notion that the deviation proportion is more than 80%, the big part has been chosen [4]. This is followed by data mining utilizing the chosen primary correlation matrix.

3 Data mining
Using a large, fragmented, chaotic, ambiguous and unpredictable input collection, data mining is a methodology for uncovering formerly undiscovered knowledge. There are two main types of information retrieval: traditional and non-traditional data mining. Multifunctional data collecting has been the subject of investigation aimed at extending traditional techniques to usefulness of information processing. For the research, a strong underpinning was provided by these experiments [5-6]. As a result of the fundamental examination of operational data elements, clusters were identified and represented in Table II. It is necessary to categorize each sample based on its distinct characteristics. Constant recurrence, as well as relatively obvious class differences, place instances with similar features and constraints in the same category.

On institutional campuses, we employ cluster analytics to identify undergraduates who use the university campus network in different ways. This allows us to properly identify how undergraduates utilize institutional campus networks and help them locate a bunch of individuals who are more worried about the network's broader utilization. It may be used to collect data that can be used to enhance the academic and reliable undergraduate administration. Modern statistical modeling approaches most commonly used are K-means and expectation maximization, respectively. For large data sets K-mean clustering approach is a fundamental and easy method, although K-validations and other techniques can be utilized. In the case of categorical information collected, there is no need to define the cluster centers in advance, and a progressive class correlation may be discovered, although this technique is only suitable for limited info. There is a tradeoff between performance and power when dealing with large data sets. As a result, the KMA was used in this work. KMA requires first determining the optimal number of groups to use Using the Hubert Index and Cross-Validation as 4 groups, the optimal number of segments was determined in this investigation, as seen in figure 1. A major part of the solution is built on procedures from the Nbclust library and R-Facto.
4. Results

4.1 Outcomes from categorization and essential data on the different students groups

To categorize participants, K-means uses 25 variables provided in this investigation as entry level input. Test participants were divided into four groups, with 110 in the initial group, 823 in the next level and 2606 in the middle and 210 in the last group.

4.2 Determinants of institutional network utilization by varied undergraduate groups

There are no differences between day time, night time, week end and evenings web usage for classes 1, 2 and 4 which is observed in figure 2. It's safe to say that the length of hours invested on the internet has not changed. It can be observed in figure 3 that total class 3 traffic, both during the day time and at night time, is considerably higher than classes 1, 2, and 4, and that the web traffic stats of the class 1 undergraduate grouping are the least. A look at figure 4 reveals a progressive increase in the stream/duration proportions for Categories 1 through 4. The 25-month stream/duration proportion for distinct undergraduate classes shown in figure 5 revealed that after the final year of college, the stream/duration proportion for each classification is dramatically improved. Class 4 and Category 3 congestion duration ratios rose significantly just after third year of college and were much higher than in the other categories.

Additional examination of the stream/duration ratio of Web traffic in four student categories throughout the day time and at night time. As a result of this, the stream/duration ratio of various varieties

![Fig. 1. Identifying the optimum clusters](image)

![Table II. The Essential data of the Varied Students Group](table)

| Category | Basic Information |
|----------|------------------|
|          | Number | Male | Percentage of Male | Female | Percentage of Female |
| 1        | 230    | 58   | 25.2%              | 172    | 74.8%               |
| 2        | 958    | 795  | 83.0%              | 163    | 17.0%               |
| 3        | 1707   | 1362 | 79.8%              | 345    | 20.2%               |
| 4        | 350    | 251  | 71.7%              | 99     | 28.3%               |
of pupils is directly comparable across the 1-25 months period (as shown in Figure 6-a). Stream-to-duration ratios for classes 3 and 4 are significantly larger, and Category 4 has been significantly better than the other (as represented in Figure 6-b). It can be shown in Figure 7-a that the aggregate Web traffic numbers and norms for solo login details are consistent amongst the four undergraduate categories in the freshmen and senior levels, indicating that various individuals in both the freshmen and senior levels. A substantial rise in the recurrence proportion of class 4 congestion has occurred during the junior year. This class has remained at the top of several classifications throughout the junior and senior. According to Figure 7-b, students in the initial grade have lesser achievement levels than their peers during the first 3 months of education, and their achievement levels are slightly better in the final year of school.

The connection information shows the characteristics of the use of various university networks by individual participants. When combined with other behavioral information from the university, it provides for a comprehensive understanding of the characteristics of distinct types of students, as well as the impact of numerous kinds of university network user attitudes. In addition to productivity information, campuses identification utilization statistics, total book lending, and health testing were incorporated in the college's statistical data.

4.3 Child behavior statistics for other school groups

University students' college campus networks are characterized by their usage characteristics, as seen by the network data. Additionally, it provides for a deeper understanding of the characteristics of different types of students in conjunction with other data acquired at schooling. It also enables a more thorough evaluation of the impact of different variants of campus network consumption behavior. In addition to academic information, campuses identification utilization records, overall textbook lending, and health testing were incorporated in the local school facts and figures. There was a decline in the mean amount of students in each group in the initial, second, third, and last grades (see Figure 8). Categories 1 through 4 have seen a decline in the outcomes of long endurance events (100 metres for males and 200 metres for girls). There was a significant difference between the efficiency of group 1 students and the other three groups in rigorous bodily fitness evaluation. Three of the four categories are quite similar to one another. When it comes to ingestion, this research suggests that the final cafeteria group consumes significantly less than the first other main divisions of breakfasts. As far as renting books is concerned, there isn't much distinction between the two sorts of students.

5. Discussion

5.1 Undergraduate University Networks in the Digital Age: Performance Improvements

Education, sociability, business and pleasure have all been significantly affected by the Web's growth as a communication technology. In the "Russian University Graduates" Online Recreation and Recreation Surveillance Survey for 2017, just 2% of undergraduate employees stated they could go digitally for more than 24 hours without their smart phones. There are 6.7% of academic undergraduates [7] who can't access the web without their cell phones. While learning and residing in university, the Internet has grown increasingly important, and cell phones have proven the main important tool for using the Web. The continuous advancement of systems administration advances has driven human progress into a time in which undergraduates today are "wherever on the Web," which has been significantly affected by conduct like learning, socialization, business and satisfaction through the Web.
Fig. 2. Online duration of undergrads

Fig. 3. Online congestion duration of undergrads

Fig. 4. Stream/duration proportions of undergrads classes 25 trimester

Fig. 5. Stream/duration proportions of 1-25 trimester

Fig. 6-a. Stream/duration proportions of morning session

Fig. 6-b. Stream/duration proportions of afternoon session
It's possible for undergraduates to utilize their smart devices and computers to rapidly access university Wi-Fi by improving the establishment of university IT infrastructure in dormitories, classrooms. In this situation, individuals may be "connected" for a long time, or even whole day. As a consequence, the reporting accuracy of the variable "Internet connected period" becomes considerably more. This report's groupings assessment showed that three sorts of information (groups 1, 2, and 4) are closely related, and the first class is slightly greater than the other 3 groups, for instance (class 3). Basically, not much has changed. To measure learners' degree of internet use, "connected duration" may accurately indicate the amount of internet usages. It is necessary to collect extra information regarding network behavior in order to analyze A participant's resource and bandwidth usage can be inferred from the amount of time they A few of the most popular internet activities are internet presence, computer communities, downloads.

If you're looking for a way to browse stationary material (message and photos) or social network activities (including Facebook), students in the third division outperform those in the first and second divisions, according to the findings of this report's supervised classification. Traffic patterns vary slightly across the four undergraduate classes. According to the first, second, third, and fourth divisions of the stream/duration proportions, the fourth factor saw a significant rise. In the stream/duration proportion, the participant's data transmission per unit time is displayed. When it comes stream/duration proportions, pupils with high-traffic activities are still more likely observed. Additionally, as seen in Fig. 8, the mean duration and median dispersion of congestion in the unique authentication connection were estimated for various groupings on university. There is a significant difference between fourth and third divisions when it comes to mean sign in duration, although the mean sign in traffic is significantly higher in division four, as seen in Figure 8. According to the earlier findings, the fourth undergraduate category preoccupies a large amount of university network and has a higher congestion pattern than some of the other groups.
Future Scope And Conclusion

As a result of this study's methodology, existing actual data processing may be analyzed, which creates awareness for forefront counselors and makes them understand the university's network of undergraduate groups. As a result, it is important to find out how undergrads spend a large amount of time at university may gain fit, healthy, happy and active knowledge as soon as possible. Academic data mining can allow teachers and undergrads administration counselors to quickly and accurately assess the health and behavior of undergraduate groups, as well as to identify students who may need help. In the data, however, there are just "impacts." "Reason" must be conveyed in reality as well. Most of the time, being a student advisor is a job for people. There are some areas of data that may be difficult to explain, such as communicating thoughts.

References

[1] Madala, S. R., Rajavarman, V. N., & Vivek, T. V. S. (2018). Analysis of Different Pattern Evaluation Procedures for Big Data Visualization in Data Analysis. In Data Engineering and Intelligent Computing (pp. 453-461). Springer, Singapore.

[2] Madala, S. R., & Rajavarman, V. N. (2018). Efficient Outline Computation for Multi View Data Visualization on Big Data. International Journal of Pure and Applied Mathematics, 119(7), 745-755.

[3] Vivek, T. V. S., Rajavarman, V. N., & Madala, S. R. (2020). Advanced graphical-based security approach to handle hard AI problems based on visual security. International Journal of Intelligent Enterprise, 7(1-3), 250-266.

[4] Romero C., Ventura S., Data mining in education[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2013, 3(1):12-27.

[5] Locantore N., Marron J., Simpson D., et al. Robust principal component analysis for functional data[J]. Test, 1999, 8(1):1–73.

[6] Yao F., Lee T. Penalized spline models for functional principal component analysis[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1):3–25.

[7] Media. 2018 Chinese College Students' Online Leisure and Entertainment Behavior Monitoring Analysis Report [EB/OL].2018-11-16.https://www.iimedia.cn/c400/62969.html.