

Mining Association Rules on Students' Profiles and Personality Types

Glenn Paul P. Gara and Francis Rey F. Padoa

Abstract— This study enables graduating high school students to choose an area of specialization in college based on their personal interest. This can be done through a Holland Code personality assessment test developed by Dr. John L. Holland [1]. Furthermore, this study is intended also to guide the school administrators in strategic planning and to enhance decision-making process. The personality assessment test will serve as a tool in generating test results that will be used for extracting patterns together with the students' profiles. The extraction can be accomplished through mining frequent patterns and association rule generation using the known apriori algorithm, which is used to uncover frequent itemsets through candidate generation [2]. The system generates frequent type of students associated to a type of personality and an area of specialization in college. Taking advantage of the said tool and method, the researchers was able to develop a system that will guide both graduating high school students and school administrators in making decisions.

Index Terms— *apriori algorithm, data mining, holland code theory, association rules, personality types*

I. INTRODUCTION

College specialization dictates the kind of career landing of a student after graduation. Nowadays, a student may choose a specialization in college based on the influences of the environment. There are 62.84% of current students who chose their specialization in college through the advice of their friends, 32.77% is through their parents' advice, 4.39% through the HEI's advice and 1.21% of the graduated students took the advice of an HEI [3]. These influences could contribute to a wrong disposition and a tendency that a student could not perform well due to lack of motivation to the program.

On the contrary, a strategic plan becomes a forefront discourse of any institution to identify threats and opportunities whether it is internal or external aspect. Higher Education Institution develops strategic plans in order to have management directions and to find its competitive advantage. Thus, the capability of an institution to take an excellent choice is very important in the face of a growing competition among other institutions. Today, the competition among higher education institutions for students has increased and so as

Manuscript received November 27, 2014; revised February 10, 2015.

Glenn Paul P. Gara, is an IAENG Member and a faculty of the Information Technology Education Program, University of the Immaculate Conception, Davao City, Philippines. e-mail: glenngara@gmail.com

Francis Rey F. Padoa, is an IAENG Member and currently an Associate Professor and Dean of the Information Technology Education Program, and also the Program Chair of the Graduate School, University of the Immaculate Conception, Davao City, Philippines. e-mail: francisreypadoa@gmail.com

the need for strategic positioning [4]. The positioning choice is a crucial decision for an institution since the position could be a central to a client's perception and choice decisions [5]. Knowledge discovery for positioning and other strategic plans can be obtained using a data analysis. Data analysis performs a significant role for supporting decisions regardless of the type of a company or institution [6].

In this study, the Holland Code Theory [1] aims to determine the possible relevant area of specialization in college based on the personality of the students. The researcher utilizes the students' profiles and test results to generate association rules. That will serve as a guide for strategic planning of a higher education institution.

II. RELATED WORK

The Holland Code test has been widely used in different systems. The Truity system [7] is an online Holland Code personality testing system that will assess the user's personality. Truity uses a 5-point likert scale in rating a particular interest. Hence, another system that is taking liberties of the Holland Code is the Personality-Testing [8] and is similar to Truity in terms of interest scaling. The test results are presented using horizontal line graph visualization and a breakdown of the six Holland Code personality types. The system 123test [9] presents the test to the users by interpreting each activity into a simple image with a brief description on top of it. Each activity serves as a test item and can be answered by ticking the check mark for yes or the x mark icon for no. The result of the test will show the dominant personality of a user accompanied with a brief description and its associated list of careers.

On the other hand, the study of association rule mining is applied several times when discovering frequent patterns. Particularly, the apriori algorithm which is known in mining frequent itemsets through candidate generation [2], [10]. Oladipupo and Oyelade [11] discovered knowledge on students' result repository to identify the failure patterns by applying the said method. The system generated patterns using the apriori algorithm in a table and a graphical representation, which were analyzed, and unveils that there is more to students' failure than the students' ability. This analysis enables academic planners from different higher education institutions to enhance its decision-making process in curriculum structure and modification to improve the academic performance of the students. Similarly, the study in mining association rules on students' assessment data [12], uses the apriori algorithm to discover association rules from the post graduate students' assessment records. The rule discovery enables academic managers and curriculum planners to redesign a

curriculum, change teaching and assessment methodologies and alter the time table timeslot in order for the students to be fully armed with the subject technicalities and to acquire desirable performance in the post graduate level. The study employs TANAGRA in rule discovery, a complimentary data mining software developed for research and academic purposes. Furthermore, association rule mining was applied in a system developed by Angeline (2013) to analyze students' performance [13]. The analysis was done using apriori algorithm in order to discover a set of rules that will help to predict students' performance from the educational database. Hence, it will also help in matching the requirements of any organizations together with the students' profile in order to provide placement for the students.

In this study, the Holland Code Theory [1] was used to determine the possible relevant area of specialization in college based on the personality of the students. The purpose of the tool is to ensure that the suggested area of specializations is anchored to their personal interests.

III. ARCHITECTURAL DESIGN

The system's functionality can be beneficial to the students and HEI administrators. In figure 1, the student can take the personality assessment test and must provide personal information using the form for profiling purposes.

As soon as the student is done with the test, the system will process it to identify the student's dominant personality and its associated area of specialization in college. A centralized database will serve as a repository of the students' data such as their profiles and test results. The administrator has a privilege to exploit the general functions of the system particularly in mining frequent patterns.

In figure 2, the mining method for discovering interesting rules is presented. In preparation for pattern analysis, the system will fetch the students' data from the database and serialize it into a JSON data format. The data format will be decoded and will be submitted to the apriori algorithm in order to analyze patterns and generate association rules. The generated rules will be sent to the Google table and chart API to visualize the results.

IV. SYSTEM OVERVIEW

The system has two main functions; personality assessment test and mining association rules on students' data. This study uses the Holland Code Theory as a tool to assess the personality of the students' personality and identify its associated list of area of specialization in college. The results of the test could help the students to decide the area of specialization to be immersed in college. Moreover, the system aims to provide data for school administrators in strategic planning, specifically the students' data. However, in order for the data to be

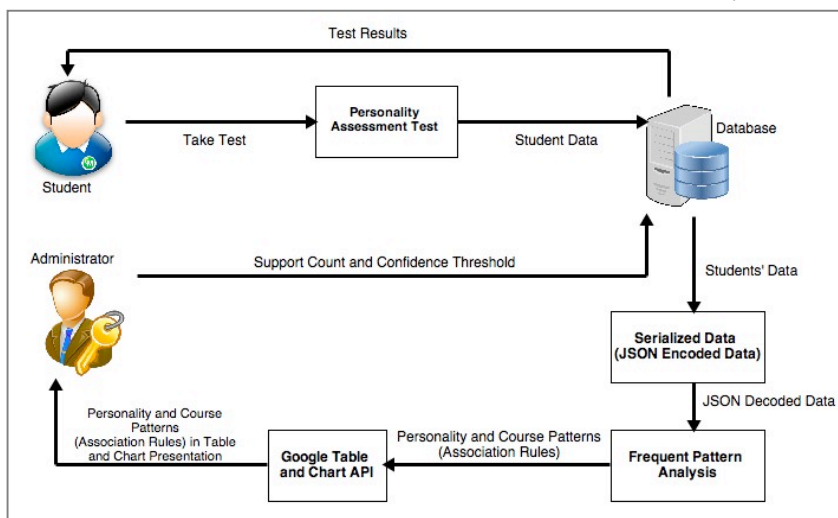


Fig 2. System architecture

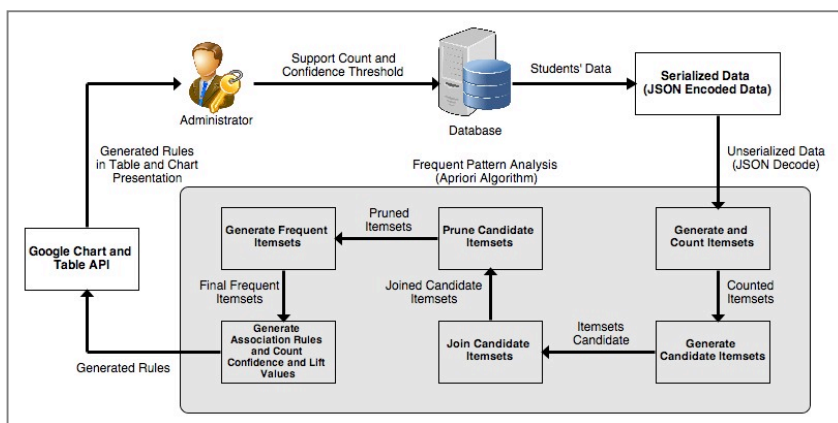


Fig 1. Mining association rules

useful and understandable in conveying information, the system utilizes a data mining technique that will generate its association rules by bringing the apriori algorithm into action.

A. Personality Assessment Test

Personality assessment test enables the graduating high school students in determining their dominant personality among the six types of Holland Code personalities, such as realistic, investigative, artistic, social, enterprising and conventional. Each of these personalities is associated to a list of area of specialization. A student will have to provide his or her profile information, specifically, the name, sex, date of birth, school and address. The profile information and the test results will be sent and stored in database that will be used by the administrator for pattern extraction and association rules generation. The system will present random questions that anchored on the personal interests of a student. Each test question is associated to a particular type of personality that can be answered by yes or no. Then, the system will process the answered test and identify the dominant personality that garnered the highest frequency among the six personality types. The system thereupon redirects the user to the result's page revealing the dominant personality type of the student and the associated list of area of specialization.

B. Mining Association Rules

The system aims to extract knowledge to guide school administrators for strategic planning and it can be attained through frequent pattern extraction and association rules generation from the students' data. The apriori algorithm enables the knowledge extraction became possible and has been put to use due to the fact that it is widely known and is considered as a standard algorithm for mining association rules [14]. The system will present to the user (administrator) the type of data available that can be filtered and processed where the setting of the minimum support count and confidence threshold take place. The minimum support count is responsible for pruning uninteresting rules while the confidence threshold measures the reliability of a rule. Both will serve as the parameters for the algorithm to extract frequent patterns and association rule generation.

1) *Data preprocessing*: Every single data that are used in pattern extraction and rule generation resides in the database. The system relies to the MySQL database for data management. The data such as sex, age, school and nationality together with the test results are pulled from the MySQL database and serialize it into a JSON format. Employing JSON enables the system to minimize database server request by sending bulk JSON encoded data.

2) *Apriori algorithm*: Apriori algorithm is widely used in different researches that centers the idea of frequent pattern and association rule mining. The algorithm was developed and proposed by Agrawal [2] in order to discover correlation between set of items in a large database. The algorithm will count the k -itemsets from the JSON encoded data and the counted k -itemsets will generate a candidate k -

itemsets. The count of the k -itemsets must be equal to or greater than to the user specified minimum support count in order to be considered as frequent itemsets. The process of joining and pruning will be applied to the frequent k -itemsets where the frequent k -itemsets generates frequent $k + 1$ itemsets. The frequent k -itemsets will be joined together and the k -itemsets that cannot satisfy the minimum support count will be pruned since they are infrequent itemsets. A k -itemsets will be considered frequent if its subsets are also frequent. The process of joining and pruning will be repeated until the frequent itemsets will become null. Thereafter, the algorithm is aborted and the association rules generation begins. The confidence level of the rule must satisfy the user specified confidence threshold. The confidence level of the rules can be calculated using the equation:

$$confidence(A \rightarrow B) = \frac{s(A \cup B)}{s(A)} \quad (1)$$

The association rules of the frequent itemsets can only be accepted if the confidence level of the rule satisfies or exceeded the user specified confidence threshold.

3) *Correlation analysis using lift*: At some point, a rule could be a misleading rule where an antecedent rule does not have any relationship to the consequent rule although it satisfies the confidence threshold, under those circumstances, the need of correlation analysis is significant. A correlation measure called lift is used to determine the correlation of a rule. It measures the interestingness of a particular rule. The lift can calculated using the equation:

$$lift(A \rightarrow B) = \frac{s(A \cup B)}{s(A) \times s(B)} \quad (2)$$

Negative correlation of a rule occurs when the resulting lift value of is less than one. It indicates that the occurrence of the antecedent rule, more likely, leads to the absence of the consequent rule. A rule is independent rule if the resulting lift value is equal to one or there is no such correlation between the antecedent and the consequent rule. Moreover, if the resulting lift value of a rule is greater than one, it follows that the antecedent and consequent rules are positively correlated. It denotes that the occurrence of the antecedent rule implies the occurrence of the consequent rule.

4) *Rules visualization*: The unpruned rules that satisfy the confidence and lift value (positively correlated) are automatically presented through the Google charts tool. The support of a rule will be supplied to the Google charts for visualization purposes. The support of the unpruned rules can be obtained using the equation:

$$support(A \rightarrow B) = \frac{s(A \cup B)}{N} \quad (3)$$

The support of a rule can be calculated by combining the support count of the antecedent and consequent rule over the total number of transactions on the dataset.

V. SIMULATION AND RESULTS

The two main functions of the system have been tested and the following sections unveiled the system results. The graduating high school students of St. Augustine International School, Davao City, Philippines were selected to test the system. In addition, the system found interesting correlations using the proposed method.

A. Students' assessment test

Figure 3 shows a profiling form where students must provide their personal information beforehand. The form will act as a conduit to collect the students' profiles. The user must fill out all the necessary fields in order to proceed to the assessment test.

In figure 4, the test item refers to a possible interest of a student. The user has to options to answer the test time, whether yes or a no. Each test item is associated to a specific type of personality of Holland Code.

Figure 5 shows the page where the system redirected the users to view the test results. The simulation revealed that the student who took the test is a Realistic type of person and

presents the list of area of specialization in college that is associated to the student's personality.

B. Mining interesting association rules

The students' data were analyzed through mining interesting association rules and discovering correlations. Figure 6 shows the admin panel where data analysis menu can be located. Data analysis is where the administrator can filter, assign minimum thresholds and process the data using the algorithm. Figure 7 shows how the data can be filtered in the data analysis page before patterns can be extracted with the algorithm. The user must set the minimum support count and confidence thresholds as a prerequisite for mining association rules. All types of data were checked in order to have a high probability to discover interesting rules. Since age is a quantitative value, it is good to use a discretization method to allow the user to transform the quantitative value into intervals in order to avoid uninteresting and redundant rules [10]. In figure 7, a 50% minimum support count and 80% confidence threshold are set. It is quite subtle when setting a minimum threshold due to the fact that it may lead to a thousand of pattern generation, whereas too huge threshold may lead to a no patterns at all [15].

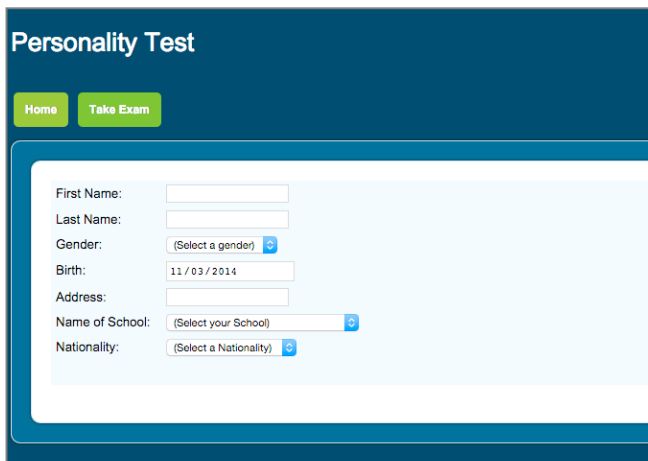


Fig 3. Profiling form

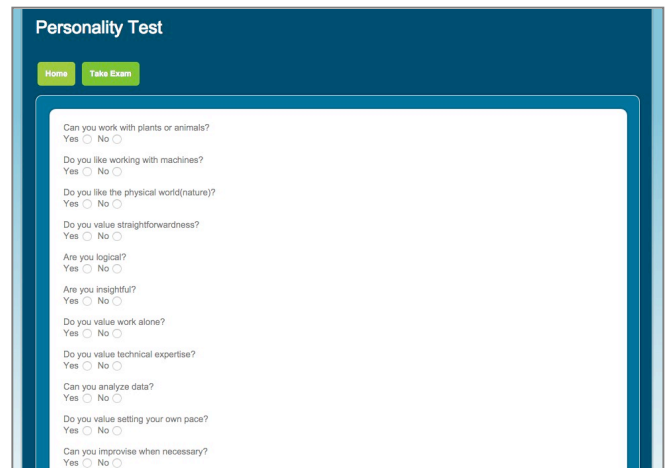


Fig 4. Assessment test

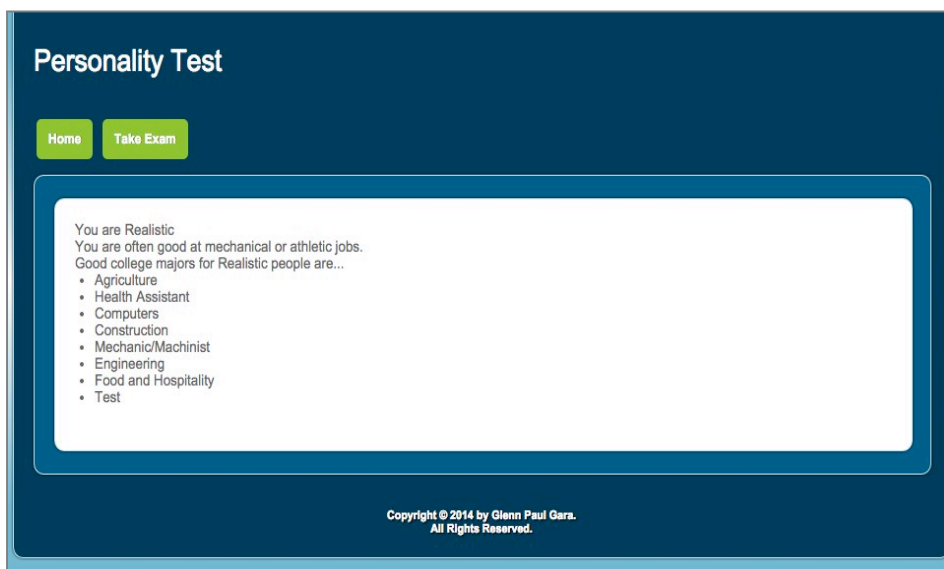


Fig 5. Sample test result

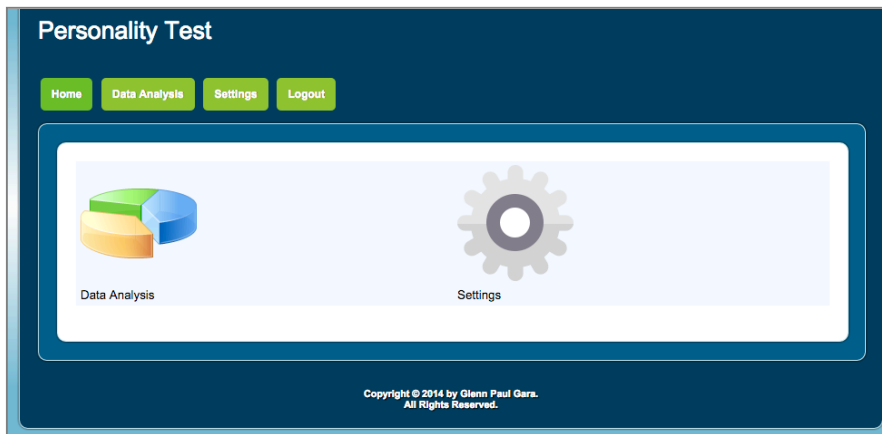


Fig 6. Administrator panel

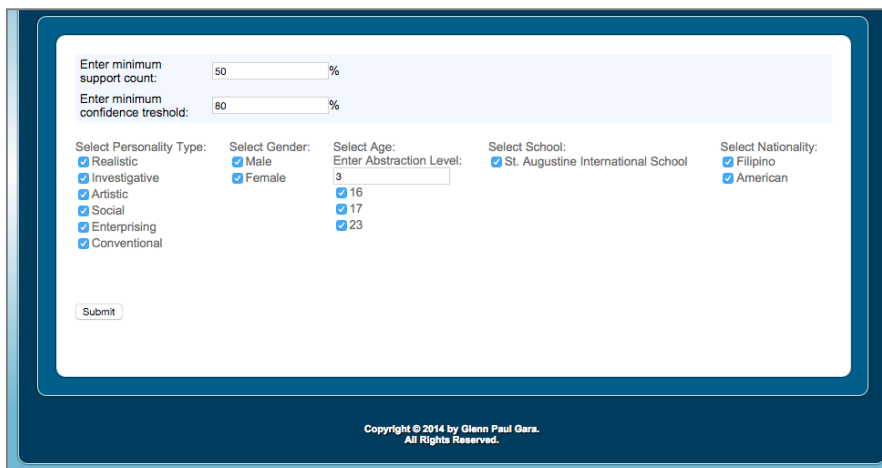


Fig 7. Data filtering and processing

Rules	Confidence	Support Count	Lift
Personality("Realistic")=>Gender("F")	81.82%	58.06%	1.01
Personality("Realistic")=>Gender("F")*School("St. Augustine International School")	81.82%	58.06%	1.01
Personality("Realistic")=>Age("16-17")	86.36%	61.29%	1.03
Personality("Realistic")=>Age("16-17")*School("St. Augustine International School")	86.36%	61.29%	1.03
Personality("Realistic")=>Age("16-17")*Nationality("Filipino")	81.82%	58.06%	1.01
Personality("Realistic")=>Age("16-17")*School("St. Augustine International School")*Nationality("Filipino")	81.82%	58.06%	1.01

Fig 8. Generated association rules

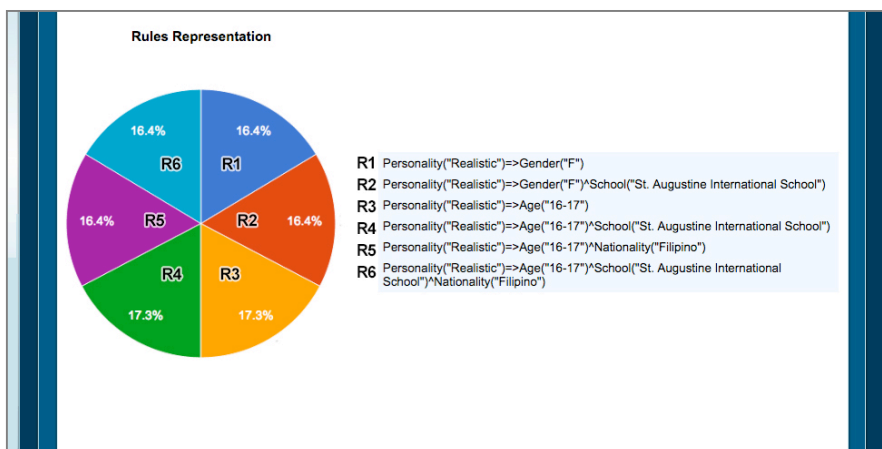


Fig 9. Results chart view

With this, setting this type of amount of thresholds allows the user to discover an interesting rules as seen in figure 8.

The system generated 6 association rules derived from the user-specified support count and confidence threshold. In figure 8, it is evident that the 3rd and 4th rules are interesting since these are the rules that generated high percentage of confidence and support count. This explains that realistic students are from ages 16-17 years old studying at St. Augustine International School. These rules are supported with a lift value of 1.03, which is greater than one indicating that the antecedent rules are positively correlated to the consequent rules. The occurrence of the antecedent rule implies the occurrence of the consequent rule. This knowledge enables higher education institution to strategically position their offered courses associated with the realistic personality to the students that ages 16-17 years old from St. Augustine International School. Thus, the market focus of the courses associated to the realistic personality must focus on these type of students. On the other hand, the rules are presented graphically using a Google pie chart as seen in figure 9 in order to visualize the support count of each rule.

VI. CONCLUSION

The study proposes a method that could guide graduating high school students and HEI administrators to enhance decision-making process. With the aid of the Holland Code developed by Dr. John L. Holland [1] the study enables the student to determine the list of areas of specialization in college based on their personal interest. The system generated results such as personality type and its associated area of specializations. In addition, the system was able to unveil interesting rules from the students' data using the apriori algorithm [2], a known method for market basket analysis. The extracted rules can be helpful for the HEI administrators to obtain position strategically, particularly to have a market focus. The results could also help them to discover the type of students that are associated to the courses that they are not currently offering. This could guide them also to decide whenever an institution is planning to offer new courses. This is to conclude that the proposed methods are feasible to support decisions to the identified beneficiary of this study.

REFERENCES

- [1] J. L. Holland, "Exploring careers with a typology: What we have learned and some new directions." *American Psychologist*, vol. 51, no. 4, p. 397, 1996.
- [2] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *ACM SIGMOD Record*, vol. 22, no. 2. ACM, 1993, pp. 207–216.
- [3] K. Ramakrishnan and N. Yasin, "Employment issues among malaysian information and communication technology (ict) graduates: A case study," *African Journal of Business Management*, vol. 6, no. 16, 2012.
- [4] P. Kotler, *Analysis, planning, implementation and control*. Prentice Hall International, 1994.
- [5] L. J. Harrison-Walker, "Strategic positioning in higher education." *Academy of Educational Leadership Journal*, vol. 13, no. 1, pp. 103– 111, 2009.
- [6] M. Goyal and R. Vohra, "Applications of data mining in higher education," *International journal of computer science*, vol. 9, no. 2, p. 113, 2012.
- [7] T. P. LLC, "Truity," 2012. [Online]. Available: truity.com
- [8] (2011) Personality-testing. [Online]. Available: personality-testing.info
- [9] (2014) 123test. [Online]. Available: 123test.com

- [10] J. Han and M. Kamber, *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann, 2006.
- [11] O. Oladipupo and O. Oyelade, "Knowledge discovery from students' result repository: association rule mining approach," *International Journal of Computer Science and Security*, vol. 4, no. 2, pp. 199–207, 2010.
- [12] V. Kumar and A. Chadha, "Mining association rules in student's assessment data," *International Journal of Computer Science Issues*, vol. 9, no. 5, pp. 211–216, 2012.
- [13] D. M. D. Angeline, "Association rule generation for student performance analysis using apriori algorithm," *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)*, vol. 1, no. 1, pp. p12– 16, 2013.
- [14] S. Mutter, M. Hall, and E. Frank, "Using classification to evaluate the output of confidence-based association rule mining," in *AI 2004: Advances in Artificial Intelligence*. Springer, 2005, pp. 538–549.
- [15] J. Han, J. Wang, Y. Lu, and P. Tzvetkov, "Mining top-k frequent closed patterns without minimum support," in *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, 2002, pp. 211–218.