Development of College Completion Model Based on K-means Clustering Algorithm

Allen M. Paz, *Member*, *IACSIT*, Bobby D. Gerardo, and Bartolome T. Tanguilig III

***Abstract*—The amount of data stored in educational databases is rapidly increasing because of the increase in awareness and application of information technology in the field of higher education. What can be done with these databases is to mine the hidden knowledge in it. This paper is designed to present and justify the capabilities of data mining. The main contribution of this paper is the development of college completion model based on k-means clustering algorithm. The data stored in the Student Information and Accounting System from 2009 to 2013 was used to perform an analysis of study outcome taking into consideration not to include in the final result any identifying information to protect their privacy. The results showed that majority of the students belong to the cluster which needs intervention. The dataset used can be improved by including data of students currently enrolled. The result obtained can be used as a decision support tool. The WEKA software was used to build the college completion model using k-means clustering.**

***Index Terms*—Database, college completion, k-means, clustering, data mining, algorithm.**

I. Introduction

Analysts of the knowledge society or knowledge economy characterize the university not just as a generator of knowledge, an educator of young minds and a transmitter of culture but also as a major agent of economic growth. It is both a research and development laboratory and a mechanism through which the nation builds its human capital to enable it to actively participate in the global economy. Hence, it is imperative for education to be shaped with in accordance to the exact needs of the industry [1].

Today higher education institutions are facing the problem of student retention which is related to college completion rates. Colleges with higher freshmen retention rate tend to have higher graduation rate within four years. Since freshmen were the most vulnerable to low student retention at all higher education institutions, early identification of vulnerable students who are prone to drop their courses is crucial for the success of any retention strategy. This would allow education institutions to undertake timely and proactive measures. Early identification of at-risk students can be the recipient of academic and administrative support to increase their chance of staying in the course and eventually complete the program.

The ability to discover hidden information from university databases particularly on enrolment data is very important in an educational institution. Being able to monitor the progress of student's academic performance is a critical issue to the academic community of higher learning. It is a long term goal of higher educational institutions to increase retention of their students. College completion is significant for students, academic and administrative staff. The importance of this issue for students is obvious: graduates are more likely to find decent jobs and earn more than those who dropped out.

With the help of data mining which is an essential process where intelligent methods are applied in order to extract data patterns, it is possible to discover the key characteristics from the students' records and possibly use those characteristics for future prediction. K-means clustering technique was employed in order to discover pattern.

The students will be the first beneficiary of any improvement on the present policies. Faculty members and advisers will be properly informed of the status of their students. This study will provide the community about the factors affecting the college completion giving them an idea on the value of the grades in high school and the scores in the college admission test as two of the requirements for admission. Once admitted, performance in the freshman year is also a determining factor in their desire to have a college diploma to have a better job and eventually better lives. This study will also help the school in providing better educational services from the time they registered for the first time until their last semester of stay in the university to complete their degree.

The only available data about the students in the university's database is the information they supplied in their enrolment form. It is a challenge on the part of the university administrators and academic planners to update records of students with relevant information that will aid in any academic related decision that may be needed in the future.

A. Research Objectives

The main objective of this study is to explore the enrolment data that may impact the study outcome of students. Specifically, the enrolment data were used to achieve the following objectives:

1) To build college completion model based on k-means clustering data mining technique on the basis of identified attributes;
2) To discover the overall distribution pattern and correlation among data attributes; and
3) To determine the significant attribute that contributed to the college completion model.

Manuscript received November 30, 2013; revised February 26, 2014.

A. M. Paz is with the College of Development Communication and Arts & Sciences, Department of Information and Communications Technology, Isabela State University, Philippines (e-mail: allenmpaz@gmail.com).

B. D. Gerardo is with the Office of the Vice President for Administration and Finance, West Visayas State University, Philippines (e-mail: bgerardo@wvsu.edu.ph).

B. T. Tanguilig III is with the College of Information Technology Education, Technological Institute of the Philippines (e-mail: bttanguilig_3@yahoo.com).

II. Background and Related Work

Reference [2] conducted a study on student factors affecting the success or failure of the Teachers Training Programs and the findings showed that majority of the students of the college were female, with average academic performance in high school, belong to the low income group, with neutral attitude towards teaching and mostly from the public schools. Thirty two percent (32%) of the students did not perform well in college as when they were in high school although their mean academic performance in college was still average. The correlation coefficients showed that while the Bachelor in Secondary Education (BSEd) males tend to perform better than the females, the females tend to perform better among the Bachelor in Elementary Education (BEEd) students. In both courses, the students coming from the private schools tend to perform better and the students with high academic performance in high school also tend to have higher academic performance in college. Only the relationship between their academic performance in high school and in college, however, was found to be significant. This means that the significant factors affecting their performance in college are school origin and grade in high school.

Reference [3] examined degree completion among college students using Astin's student typology framework. The study was complex and yielded a mix of statistically significant findings. However, four findings emerged from his study. First, degree completers are more likely to earn better high school grades than dropouts. Second, middle and high-income students are more likely to graduate from college than low-income students. Third, for status striver type of students, other (non-academic) background variables predict college academic performance in terms of Grade Point Average (GPA) and total college credits. Fourth, for Social Activist type of students, other (non-academic) background variables predict grades earned in college.

The study conducted by [4] examined the validity of High-school grades in predicting student success beyond the freshman year. The results showed that high-school grade point average (HSGPA) is consistently the best predictor not only of freshman grades in college, but of four-year college outcome as well. The study tracked four-year college outcomes, including cumulative college grades and graduation, for the same sample in order to examine the relative contribution of high school record and standardized tests in predicting longer term college performance. Key findings showed that HSGPA is consistently the strongest predictor of four-year college outcomes for all academic disciplines. The predictive weight associated with HSGPA increases after the freshmen year, accounting for a greater proportion of variance in cumulative fourth-year than first-year college grades. Other factors such as standardized test, school academic performance index, socio-economic status and parents education were considered by only to concede to HSGPA as a valid factor for predicting success beyond freshman year.

A model was developed using a structural equation modeling to explain college completion of undergraduate students [5]. The independent variables were perceived institutional support, academic self efficacy, institutional commitment, classroom learning environment and social support. The conclusion reached from the analysis is that the learning environment is a moderately powerful but indirect influence on student college completion intention. Social support and perceived institutional support contribute to a student's intention to complete college. Academic self-efficacy also plays a smaller yet significant role in student's college completion intention.

Reference [6] conducted a study on the student performance by means of Bayesian classification method on 17 attributes, it was found that the factors like students' grade in senior secondary exam, living location, medium of teaching, mother's qualification, students other habit, family annual income and student's family status were highly correlated with the student academic performance.

The application of Data Mining in the education sector was explored by [7]. The study takes the performance of students in their examination and their presence in the classroom and finds a relation in them. The observed relation helps in identifying the group of students where the extra attentions are required. The study was carried out using K-means method of cluster analysis.

A study conducted by [8] revealed that preadmission scholastic assessment test (SAT) scores and high school record are significant predictors of graduation. The correlations observed were moderate and lower than the correlations of admission credentials with cumulative GPA. Other predictors and criteria of success which are non-academic and which clearly influence persistence in college are financial status, health and student personality.

A case study was presented on educational data mining to identify up to what extent the enrolment data can be used to predict student's success [9]. The algorithms *Chi*-squared Automatic Interaction Detector (CHAID) and Classification and Regression Tree (CART) were applied on student enrolment data of information system students of open polytechnic of New Zealand to get two decision trees classifying successful and unsuccessful students. The accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively.

An intelligent student advisory framework in the educational domain was developed by [10]. They classified the students into the suitable department using C4.5 algorithm. They also clustered the students into groups as per the suitable education tracks using k-means algorithm. They combined the results that came out from classification and clustering operations to predict more results. A case study was presented to prove the efficiency of the proposed framework. Students data collected from Cairo Higher Institute for Engineering, Computer Science and Management during the period from 2000 to 2012 were used and the results proved the effectiveness of the proposed intelligent framework.

Reference [11] used a data mining approach to differentiate the predictors of retention among freshmen enrolled at Arizona State University. Using the classification tree based on an entropy tree-splitting criterion they concluded that 'cumulated earned hours' was the most important factor contributing to retention. Gender and ethnic origin were not identified as significant.

Reference [12] conducted a study to analyze students'

results based on cluster analysis and used standard statistical algorithms to arrange their scores according to their level of performance. K-means clustering algorithm was implemented. The model created was an improvement of the limitation of existing methods developed by Omelehin using fuzzy logic.

Reference [13] in their study presented a hybrid procedure based on decision tree of data mining method and data clustering which will enable academicians to predict student's GPA and based on that, instructors can take a necessary step to improve student academic performance.

III. Work Done/Contribution

A. Framework of the Study

The framework of the study was based on the Knowledge Discovery Process (KDP) illustrated by [14]. The KDP figure was modified to suit the objectives of the study. The modified version was presented on Fig. 1 following the steps from **preprocessing** wherein noisy and irrelevant data were removed, **selection and transformation** where data relevant to the analysis task were retrieved from the database and further transformed or consolidated into forms appropriate for mining, **data mining** where k-means clustering were applied in order to extract data patterns, **interpretation and evaluation** where the truly interesting patterns representing knowledge based were identified and **knowledge presentation** where visualization and knowledge presentation techniques were used to present the mined knowledge to the user.

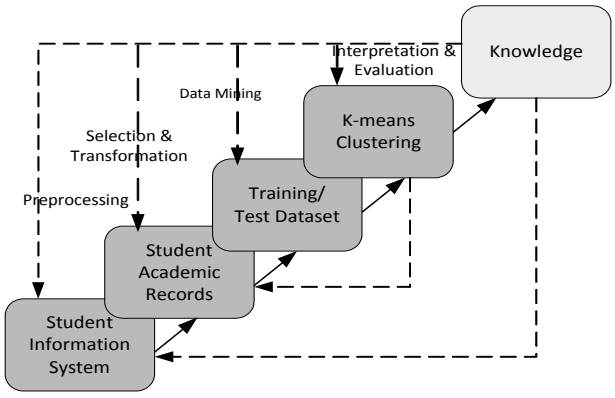

Fig. 1. The steps of extracting knowledge from data.

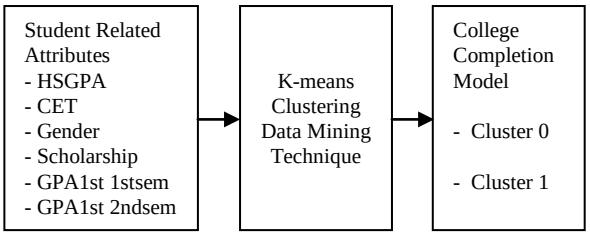

Fig. 2. College completion model framework using K-means clustering algorithm technique.

B. College Completion Model Process Framework

Data mining is just a part of the whole framework of the study. Fig. 2 shows the college completion process framework as its major components used in this study.

The information stored in the Student Information and Accounting System were analysed to be able to extract appropriate dataset for the study. The dataset was produced after data pre-processing. This served as input to the data mining tool for the application of the selected k-means clustering algorithms to develop the college completion model. Two clusters were produced after the process.

The model was evaluated based from their accuracy consistent with the results obtained from training dataset. Cluster number represents groups of student related or similar with each other. The knowledge discovered can then be used for decision making.

C. Methodology/Data Mining Process

The data mining tool used in this study is Weka which offers different data mining techniques for various kinds of data. The Weka Knowledge Explorer is an easy to use graphical user interface that harnesses the power of the Weka software. The major Weka packages are Filters, Classifiers, Clusters, Associations, Attribute Selection and Visualization tool, which allows datasets and the predictions of Classifiers and Clusters to be visualized in two or three dimensions. The workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling together with graphical user interfaces for easy access to this functionality. Weka was primarily designed as a tool for analysing data from agricultural domains. It is now used in many different application areas, in particular for educational purposes and research [15].

K-means clustering technique was selected to analyze the dataset extracted from the Student Information and Accounting System of the Isabela State University. Fig. 3 shows the process of how k-means clustering work. A clustering algorithm attempts to find natural groups of components (or data) based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster.

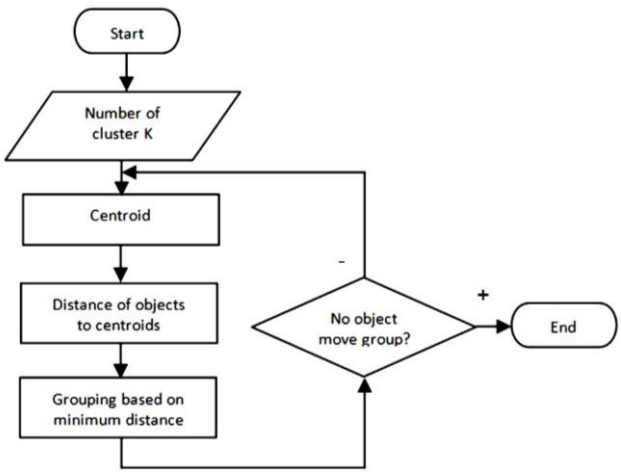

Fig. 3. K-means clustering process flowchart.

Fig. 4 shows that 392 out of 1053 students or 37.23% completed their enrolled program on time. The rest of the students are either dropped or still enrolled to complete the program.

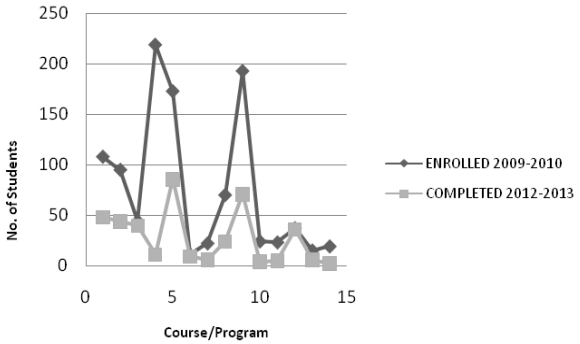

Fig. 4. Enrolled and completed indicator.

A total of 164 academic records of freshmen students enrolled in various programs in the university in the school year 2009-2010 were taken as a sample from a total population of 1053 freshmen students. Students from other institutions who transferred to the university with earned units were not included in the study. Description of variables and their data types were presented in Table I.

The domain values of the attributes used were defined as follows:

HSGPA-High School Grade Point Average. It is the general weighted average of the student in the last year in high school.

CET-College Entrance Test. It is a standardized test given to student who intends to enroll in the university. The examination is composed of 100 multiple test items.

GENDER-Student's gender. It is the category of student whether male or female. M represents male while F represents Female.

SCHOLAR-Scholarship grant. It is a field with yes or a no value. Y represents student with scholarship and N represents student without scholarship.

GPA11ST-Grade Point Average in first year first semester. This is the average grade of student while in first year first semester. The value range from 1.0 to 5.0 where 1.0 is the highest grade a student can get and 5.0 as the lowest grade.

GPA12ND-Grade Point Average in first year second semester. This is the average grade of student while in first year second semester. The value range from 1.0 to 5.0 where 1.0 is the highest grade a student can get and 5.0 as the lowest grade.

TABLE I: STUDENT RELATED ATTRIBUTES AND DATA TYPES

Variable	Description	Data Type
HSGPA	High School Grade Point Average. This is the general average obtained in their last year in high school	Numeric
CET	College Entrance Test. This is the score obtained in the examination given by the university before entering college	Numeric
GENDER	Student's gender. This is the category of students whether male or female	Nominal
SCHOLAR	This is the category of student whether the student is enjoying scholarship or not	Nominal
GPA11st	Grade point average in first year first semester	Numeric
GPA12nd	Grade point average in first year second semester	Numeric

IV. RESULTS AND DISCUSSIONS

The objective of the study is to build model for college completion using K-means clustering technique on the basis of identified attributes which were HSGPA, CET, GENDER, SCHOLAR, GPA11ST, and GPA12ND. The result is shown below:

Fig. 5 shows the output of k-means clustering algorithm when executed. Two clusters were formed after 7 iterations. Cluster 0 contains 85 instances or 52% while Cluster 1 contains 79 instances or 48%. This analysis showed that in Cluster 0, 85 out of 164 students had a HSGPA score of approximately 83, CET scores of approximately 39, mostly female, mostly scholars, with a GPA during their first year first semester of approximately 2.92 and approximately 3.34 during second semester of the same year level. The rest of the students belong to Cluster 1 with approximately 87.46 HSGPA, approximately 51.16 scores in CET, mostly female, mostly scholars, with a GPA during their first year first semester of approximately 2.28 and approximately 2.58 during second semester of the same year level.

```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 99.64107990750725
Missing values globally replaced with mean/mode

Cluster centroids:
                          Cluster#
Attribute     Full Data          0          1
                 (164)       (85)       (79)
=============================================
HSGPA          85.1366    82.9782    87.459
CET            44.8598         39   51.1646
GENDER               F          F          F
SCHOLAR              Y          Y          Y
GPA11ST          2.611     2.9209     2.2776
GPA12ND          2.964     3.3375      2.562


Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       85 ( 52%)
1       79 ( 48%)
```

Fig. 5. Model and evaluation on training set.

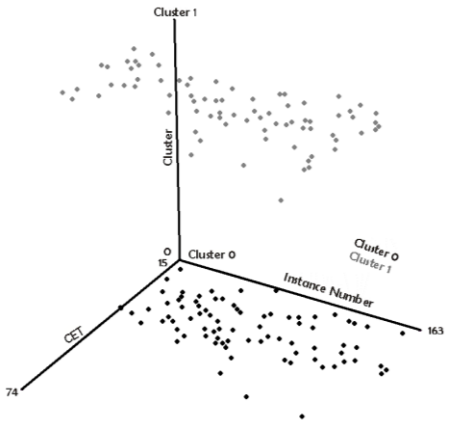

Fig. 6. Graphical representation of clustered instances.

Fig. 6 shows a graphical representation of clustered instances. Cluster 0 is in the lower region of the axes which is compose of 85 students while Cluster 1 is on the upper region of the axes which is compose of 79 students.

The results showed that students from Cluster 1 are more likely to complete college on time than those students in Cluster 0. Students in Cluster 0 are considered at-risk students. They are the students who are more likely to drop or stay longer in the university to finish college.

The statistical tool SPSS was utilized to discover the overall distribution pattern and correlation among data attributes. Fig. 7 shows the correlation matrix

The result showed that the attributes CET and HSGPA, GPA11ST and GPA12ND were strongly correlated.

It shows that the College Entrance Test got the highest correlation coefficient among the 6 variables which implied that CET is strongly correlated and highly significant.

Correlations

		HSGPA	CET	GENDER	SCHOLAR	GPA11ST	GPA12ND
HSGPA	Pearson Correlation	1	.481**	-.217**	.063	-.470**	-.431**
	Sig. (2-tailed)		.000	.005	.425	.000	.000
	N	164	164	164	164	164	164
CET	Pearson Correlation	.481**	1	-.106	.106	-.383**	-.301**
	Sig. (2-tailed)	.000		.176	.176	.000	.000
	N	164	164	164	164	164	164
GENDER	Pearson Correlation	-.217**	-.106	1	.018	-.033	-.060
	Sig. (2-tailed)	.005	.176		.821	.675	.446
	N	164	164	164	164	164	164
SCHOLAR	Pearson Correlation	.063	.106	.018	1	.003	-.069
	Sig. (2-tailed)	.425	.176	.821		.971	.377
	N	164	164	164	164	164	164
GPA11ST	Pearson Correlation	-.470**	-.383**	-.033	.003	1	.755**
	Sig. (2-tailed)	.000	.000	.675	.971		.000
	N	164	164	164	164	164	164
GPA12ND	Pearson Correlation	-.431**	-.301**	-.060	-.069	.755**	1
	Sig. (2-tailed)	.000	.000	.446	.377	.000	
	N	164	164	164	164	164	164

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 7. Correlation matrix.

V. Conclusion

The study examined the available enrolment data of students in the university's database. Based on result from k-means clustering, 52% of 164 freshmen enrolled were considered at-risk of not completing their programs on time while 48% has a greater chance of completing college. For the college completion model obtained, it can be concluded that score in the College Entrance Test is a significant factor in determining college completion as it gets a correlation coefficient of +.481. This paper is an endeavour in providing the new method of taking advantage of the available data for the improvement of educational process via data mining technology. The main idea is to come up with a college completion model which can be used to improve the decision making processes.

VI. Future Work

Future researchers may use the model to identify the existing area of research in the field of data mining in higher education. They may use additional predictor variables related to students and institution that may have effect on the retention and college completion of students. The inclusion of the records of currently enrolled students is highly recommended to monitor their progression and for early intervention for those who may be considered at-risk. The development of decision support system may also be undertaken for a more efficient monitoring and effective decision making.

Acknowledgment

The author would like to express her gratitude to Dr. Bobby D. Gerardo for the knowledge gained in Data Mining and for the conceptualization of this paper. Gratitude is also extended to Dr. Bartolome T. Tanguilig III., Dr. Lorena Rabago, and Dr. Ariel Sison for their valuable suggestions for the improvement of this paper. Special thanks to Dr. Alto for patiently answering the author's inquiry related to the study. Appreciation and gratitude is also due to Dr. Ambrose Hans G. Aggabao for his input on the statistical inquiry of the author and to the personnel of the Registrar's Office for patiently assisting the author in the data gathering process. Special appreciation is also given to Mr. Randy Macapallag for allowing the author to explore the Student Information System and extract valuable data.

References

[1] Presidential Task Force for Education. "The Philippine main education highway: towards a knowledge-based economy," Preliminary report. Manila, 2008.

[2] E. C. Toquero and M. S. Leaño, "Student factors affecting the success/failure of the TC-ISUE teacher taining programs," *Research and Development Journal*, vol. 24, January-June 2003.

[3] E. A. Miller, "Degree completion among college students and Astin's student typology framework," Virginia Polytechnic Institute and State University, November 2004.

[4] S. Geiser and M. V. Santelices, "Validity of high school grades in predicting student success beyond the freshman year: high school records vs. standardized tests as indicators of four-year college outcomes," Center of Studies in Higher Education, University of California, Berkeley, 2007.

[5] D. Thomas, "College completion intention: a structural equation model," Adventist International Institute of Advanced Studies, Silang, Cavite, Philippines, 2013.

[6] B. K. Bharadwaj and S. Pal, "Data mining: a prediction for performance improvement using classification," *International Journal of Computer Science and Information Security*, vol. 9, no. 4, pp. 136-140, 2011.

[7] S. P. Singh, B. K. Sharma, and N. K. Sharma, "Use of clustering to improve the standard of education system," *International Journal of Applied Information Systems*, vol. 1, no. 5, pp. 16-20. February 2012.

[8] N. W. Burton and L. Ramist, "Predicting success in college: SAT studies of classes graduating since 1980," College Entrance Examination Board, New York, 2001.

[9] Z. J. Kovacic, "Early prediction of student success: mining student enrollment data," in *Proc. Informing Science and IT Education Conference, 2010*.

[10] H. M. Nagy, W. M. Aly, and O. F. Hegazy, "An education data mining system for advising higher education students," *International Journal of Computer, Information Science and Engineering*, vol. 7, no. 10, 2013.

[11] C. H. Yu, S. Digangi, A. Jannasch-Pennell, W. Lo, and C. Kaprolet, "A data-mining approach to differentiate predictors of retention," presented at the EDUCAUSE Southwest Conference, Austin, Texas, USA 2007.

[12] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, "Application of K-means clustering algorithm for prediction of students' academic performance," *International Journal of Computer Science and Information Security*, vol. 7, no. 1, 2010.

[13] M. H. I. Shovon and M. Haque, "An approach of improving student's academic performance by using K-means clustering algorithm and decision tree," *International Journal of Advanced Computer Science and Applications*, vol.3, no. 8, 2012.

[14] J. Han and M. Kamber, "Data mining: concepts and techniques," Simon Fraser University, Morgan Kaufmann publishers.

[15] S. K. Yadav, B. Bharadwaj, and S. Pal, "Data mining applications: a comparative study of predicting student's performance," *International Journal of Innovative Technology and Creative Engineering*, vol. 1, no. 12.

Allen M. Paz was born in Taysan, Batangas, Philippines on October 11, 1968. She took her undergraduate program from Adamson University, Manila, Philippines, with a degree of bachelor of science in computer engineering in 1990. She finished her master's degree in information technology from the University of La Salette, Isabela, Philippines in 2004. She is currently taking her doctor in information technology at the Technological Institute of the Philippines, Quezon City.

She is presently working at the Isabela State University as a faculty member with a rank of assistant professor II. She was designated department chairman of the Information and Communications Technology Department from 2003 to 2010 in the College of Development Communication and Arts & Sciences of Isabela State University, Philippines. She is presently working on her research in data mining.

Prof. Paz is a member of International Association of Computer Science and Information Technology (IACSIT), International Association of Engineers (IAENG), Philippine Society of IT Educators (PSITE) and Isabela State University Faculty Association. She is an active member of Accrediting Agency of Chartered Colleges and Universities of the Philippines (AACCUP).

Bobby D. Gerardo is currently the vice president of Administration and Finance of West Visayas State University, Iloilo City, Philippines. His dissertation is "Discovering driving patterns using rule-based intelligent data mining agent (RiDAMA) in distributed insurance telematic systems." He has published 54 research papers in national and international journals and conferences. He is a referee of international conferences and journal publications such as *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Knowledge and Data Engineering*. He is interested in the following research fields: distributed systems, telematics systems, CORBA, data mining, web services, ubiquitos computing and mobile communications.

Dr. Gerardo is a recipient of CHED Republica Award in Natural Science Category (ICT field) in 2010. His paper entitled "SMS-based automatic billing system of household power consumption based on active experts messaging" was awarded best paper on December 2011 in Jeju, Korea. Another best paper award for his paper was "Intelligent decision support using rule-based agent for distributed telematics systems," presented at the Asia Pacific International Conference on Information Science and Technology, on December 18, 2008. An excellent paper award was given for his paper "Principal component analysis mechanism for association rule mining," on Korean Society of Internet Information's (KSII) 2004 Autumn Conference on November 5, 2004. He was given a university researcher award by West Visayas State University in 2005.

Bartolome T. Tanguilig III was born on February 24, 1970 in Baguio City, Philippines. He took his bachelor of science in computer engineering in Pamantasan ng Lungsod ng Maynila, Philippines in 1991. He finished his master degree in computer science from De la Salle University, Manila, Philippines in 1999. His doctor of philosophy in technology management was awarded by the Technological University of the Philippines, Manila in 2003.

He is currently the assistant vice president of Academic Affairs and concurrent dean of the College of Information Technology Education and Graduate Programs of the Technological Institute of the Philippines, Quezon City. His research entitled "J-master: an interactive game-based tool for teaching and learning basic java programming" was awarded the best research in the 10th National Convention for IT Education held in Ilocos Norte, Philippines in 2012. He published a research entitled "Predicting faculty development trainings and performance using rule-based classification algorithm" in Asian Journal for Computer Science and Information Technology.

Dr. Tanguilig is a member of Commission on Higher Education Technical Panel for IT Education, Board Chairman of Junior Philippine IT Researchers, member of Computing Society of the Philippines and Philippine Society of IT Educators-NCR.