Development and Assessment of Outdated Computers: A Technology Waste for Alternative using Parallel Clustering

Jeffrey John R. Yasay

Department of Computer Studies, College of Engineering and Technology
Tarlac Agricultural University, Camiling Tarlac, Philippines

jryasay@tau.edu.ph

Abstract. Technology is constantly evolving to the point that computers that are purchased then are inevitably outmoded in terms of speed and their ability to process new applications. The study aims to provide procedure and measurement in viewing the process of the parallel clustered computers via graphical representation. The idea of the development procedure has been conceptualized by the author to elevate obsolete computer for alternative use. Likert scale was used (experts and users) in assessing the system. It was found out that the development has a promising result as evident in the assessment of experts on the system's reliability (availability and stability) and the users' assessment of the system's accessibility (ease of use and flexibility). It is also noted that obsolete computers have alternative disposal technique of e-wastes. With this, the development of clustering (using the interconnectivity of a master node and slave nodes) that is reliable, accessible and with a minimal cost was conceptualized as an alternative for managing e-waste and addressing the demand of new technology in the public sectors.

Keywords: Parallel clustering, processing power, alternatives & e-waste.

1 Introduction

Nowadays, technology has become an essential part of our lives. New technology has paved the way for smartphones, faster and more powerful computers, more compact televisions and so much more. Technology has made our lives simpler, quicker, safer and more enjoyable.

Technology has truly revolutionized the way we live and the way we work. It has provided opportunities for productivity and development. It has made working more effective and efficient in general as companies continue to invest in cutting-

edge technologies.

With all the promising outcomes of technology, companies have embraced it and enjoy all the profits it could give. It has played a crucial role in companies that technology is no longer seen as cost but more of an investment. At present, various companies and industries have strategically advanced their technologies to cope with the ever-changing world.

However, as technology progresses, there were also setbacks created by them. So much of the wastes from various industries come from the technologies that are utilized in their gateways. E-wastes, or the electronic products nearing the end of their "useful life" such as computers, televisions, copiers, and fax machines are some of the challenges in the fast-paced technology development.

E-waste, also known as "a wide and growing range of electronic devices ranging from large household appliances such as refrigerators, air conditioning, cell phones, personal stereos and consumer electronics to computers that have been discarded by their users" [1], has a major effect as technology progresses. Technology has developed and progressed so fast. Rapid application development has become challenging for developers to adapt, although some are searching for alternatives that will potentially help urbanized communities develop those technology.

As we live in a world that is geographically complex and unpredictable, new business forces are generated by the rush of mega-trends, including dramatic shifts in globalization and advances in technology. For any organization to survive and prosper in such an environment, innovation is imperative.

However, innovation is no longer just for creating value to benefit individuals, organizations, or societies. Innovation's overall goal can be far more far-reaching, helping to build a smart world where people can achieve the highest possible quality of life [2].

Over the past decade, technical advances have accelerated the exponential use of multimedia tools by learners of all ages. These global trends also include the constant progression of the e-learning assessment. Evaluation is the practice of clarifying what needs to be done and relating it to what needs to be done, in order to promote the evaluation of performance and how it should be achieved [3]. In terms of speed and their ability to process new applications, computers which are then bought are ultimately outdated. When this happens, outdated computers are considered to be redundant. This also happens in sectors where computation plays a crucial role in development and achievement. As necessity dictates, there is a need to find a way in which these devices, considered redundant and worthless, can be useful in constructing computers that can meet the demands of whatever

endeavours.

A cluster consists of a series of interconnected stand-alone computers operating together as a single consolidated computing resource and is a type of parallel or distributed computer system [4]. Clustering is commonly used in a network to reduce the energy consumption and thus increase the network longevity [5]. In other terms, cluster is a series of separate and inexpensive computers, used together to provide a solution as a supercomputer.

Cluster computing provides a single general approach for designing and implementing high-performance parallel systems independent of individual hardware manufacturers and their product preferences [6]. A typical application of cluster parallel computing is to load and disperse the demand for processes by the master node to the slave nodes. The information is transmitted from the source to its respective cluster head and then to the base station in order for the selected head to bear all of the information that needs to be transmitted and route it to the intended target [7]. A commodity cluster is an array of entirely autonomous computer systems that are interconnected by an off-the-shelf networking network of commodity interconnections [8] and play a major role in redefining the supercomputing concept. As a result, high-performance high-throughput, and high-availability computing has arisen as parallel and distributed standard platforms.

With this, the development of clustering (using the interconnectivity of a master node and slave nodes) that is reliable, accessible and with a minimal cost was conceptualized as an alternative for managing e-waste in the public.

2 Build and Architecture

2.1 The parallel clustered uniform set-up

After the selection of obsolete system attachments on peripherals, cluster computers must be built.

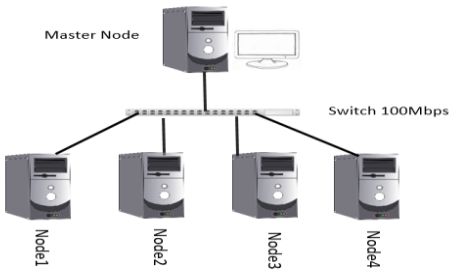

Fig. 1. Indicates the cluster clustering connectivity. The development was based on computer architecture clustered in parallel.

2.2 Production Instruments

The design of the clustered computers was based on the hardware and software needed to meet the demand of cluster computers are (a) personal computers consist of the same basic components: a CPU, memory, circuit board, storage, and input/output devices [9] (b)fast ethernet switch [10] (c)straight cable (T568A – T5668A) [11] and (d) Ubuntu ABC GNU/Linux [12].

2.2 Setup Clustering

Homogeneous computing is used to interconnect identical processor cores or units to create a high-performance device in order to use a homogeneous parallel clustering mechanism [13]. The nodes 1-4 and the master node all come in the same "Boot to Network" BIOS (basic input output system) configuration connected via T568A using Cat-5E UTP cable.

2.3 Installation (Software)

The next move is to install the program after the computers have been assembled. ABC GNU Linux (Ubuntu 9.04) [10] was used with the default kernel as a basis. Upon the installation of ABC GNU Linux (Ubuntu 9.04), gathered the information about the hardware specifications.

2.4 Specification and checking of device

Step 1: Upon determining the master node and slave node this will be the basis of hetorogenousity of the system as the specification be Processor: Intel Celeron M CPU with a CPU Speed: 2266 MHz

Step 2: Setting up of ABC GNU Linux kernel ISOLINUX3.63 Debian to the master node. Boot from the CD-ROM then choose an install mode, press enter then

follow the directions on the screen. The default language of the distro is Spanish. Changing to your preference language is necessary. After which select use entire disk to partition the hard disk, then create username and password and lastly install ABC GNU (Ubuntu 9.04)

Step 3: Setting up the slave nodes, first enter the configuration or setup of CMOS, choose halt on ALL ERROR, and finally set-up to boot from the network.

Step 4: This procedure will check the master node via Command Line Interface (CLI), master@master-desktop:~$ cat clusterhosts 192.168.0.1. Upon checking proceed to connectivity check this will test the network connectivity of Master Node, Node1, Node 2, Node 3, and Node 4, master@master-desktop:~$ cat clusterhosts 192.168.0.1 192.168.0.13 192.168.0.3 192.168.0.10 192.168.0.8.

3 Monitoring

3.1 Cluster interpretation of GANGLIA monitoring tool (GUI) [14]

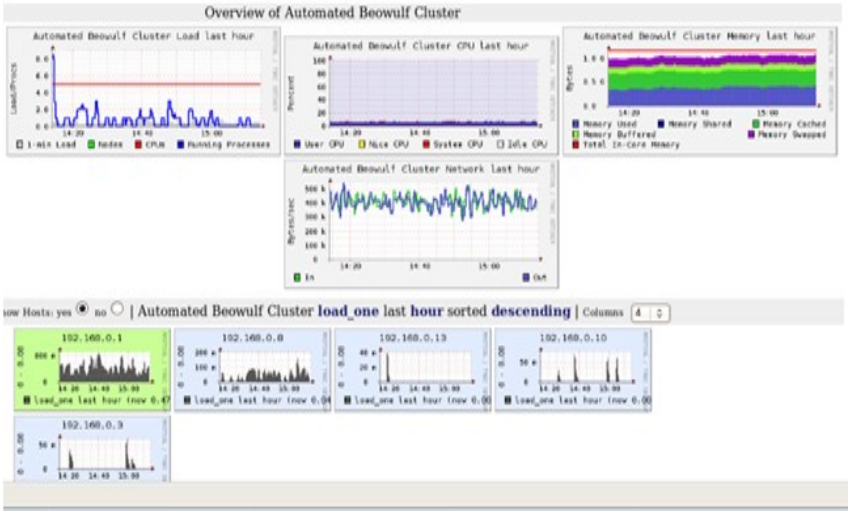

Fig 2. Overview of Automated Beowulf Cluster using Ganglia

Fig 2. Shows the device view of the cluster. A series of small graphs display the master node, and processes are used for nodes 1-4. It also indicates that the master node and nodes 1-4 work with various processes.

3.2 Performance differences of machine loaded

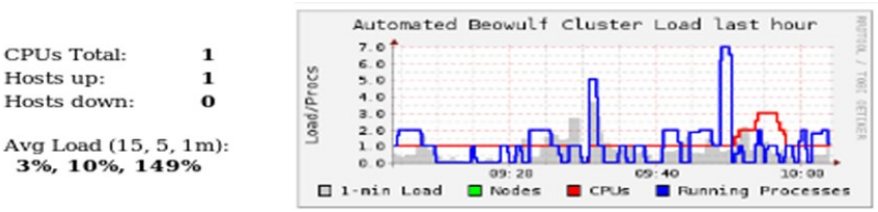

Fig 3. Performnace of Total hosts (1 CPU)

Fig 3. Displays performance representation from 1 host. It showed that the average capacity of a single CPU was 3%, 10% and 149%, showing that it is hard for a single host to process.

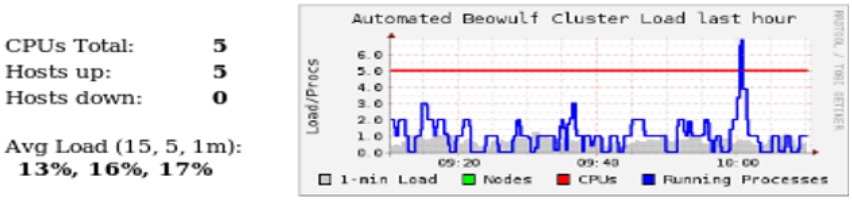

Fig 4. Performnace of Total hosts (5 CPU)

Fig 4. Shows the performance of 5 host computer. It indicates that the average load of performance is 13% , 16% and 17% which reveal that a multiple hosts process smoothly.

3.3 Network flow by graph (Master Node and Node 1)

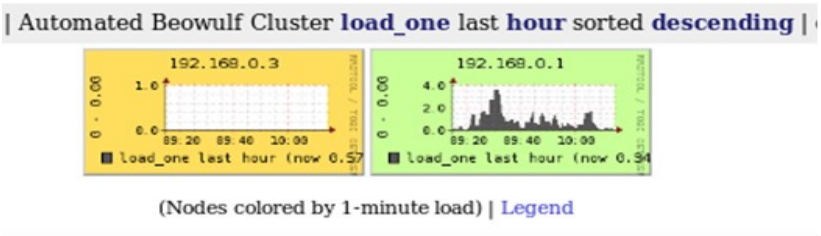

Fig 5. Master Node and Node 1

Fig 3. Reveals the master node and node 1. It ensures that the Master Node process and Node 1 process are distinct from one another. This also shows how process efficiency and relation identification are calculated.

3.4 Network movement process by graphs (Master Node and Node 1-4)

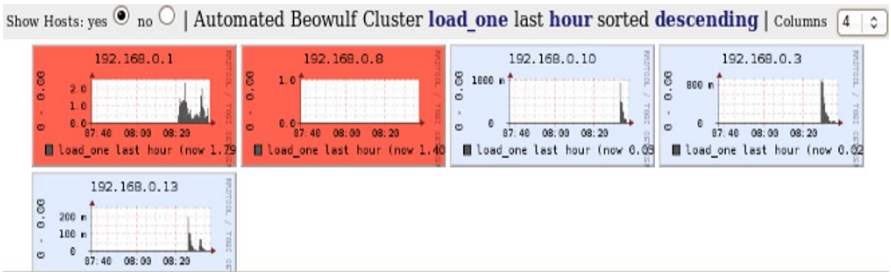

Fig 6. Process Identification of Nodes

Fig 6. Shows that the use of the CPU is 100%, it also shows that the Master Node and Node 1 used their processing power in the process distribution. It also reveals that different nodes have distinct processes.

3.5 Network movement process by graphs in Shutting down of Nodes

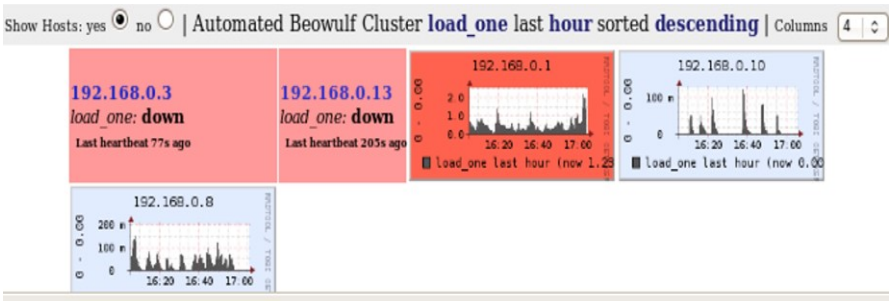

Fig 7. Node Process in Shutting Down

Fig 7. Indicates the Nodes have been successfully shut down. In the image and graph, the master node and the remaining nodes used that homogeneous parallel clustering processes are established and used.

4 Evaluation and Results

Two approaches are applied to test the homogeneous parallel clustering of alternatives for success acceptance and creation: by IT experts and by the users. The IT experts assessed the system as to Reliability with system availability and system stability [15] while the users rated the system as to accessibility with ease of use and flexibility of the system [16]. The questionnaire was based on the

Likert scale suggested by ISO 9126 [17] and used to analyze the results from scales 4.01-5.0 as excellent, 3.01-4.0 as very good, 2.01-3.0 as good, 1.01-2.0 as fair and 0-1.0 as poor with the following informative equivalents.

4.1 IT Experts

Table 1. Assessment of the System by IT Experts

Assessment Criteria	Mean	Descriptive Rating
Realiability (Composite Mean: 4.08)		
System Availability	4.50	Excellent
System Stability	3.67	Very Good

Table 1 shows the results of the evaluation based on the reliability of the system. It obtained a composite mean of 4.08.

The IT Experts evaluated the reliability of the system based on the system availability with a 4.50 mean with a descriptive rating of Excellent and system stability with a 3.67 mean with a descriptive rating of Very Good.

4.2 Assessment of Users

Table 2. Assessment of the System by Users

Assessment Criteria	Mean	Descriptive Rating
Accessibility (Composite Mean: 4.69		
Ease of Use	4.67	Excellent
Flexebility of the System	4.72	Excellent

Table 2 shows the results of users' assessment using a homogeneous parallel cluster. The users of the system were the students, IT faculty, and employees of Tarlac Agricultural University. To obtain the reliability of the evaluation, there were sixty (60) users who evaluated the system.

They evaluated the system accessibility based on ease of use with 4.67 as excellent and flexibility of the system with 4.72 as excellent. The system accessibility obtained a composite mean of 4.69 with a descriptive rating of excellent. The result indicate the uncomplicatedness of the system's operation.

5 Conclusion

The study found that the development and assessment result of the homogeneous parallel process clustering as alternatives is significant. Over the course of the review and testing, the performance of the machine was not damaged. Hence, the achievement of serviceable machines with low development costs has been established and guaranteed. Moreover, based on expert opinion and review, the use of a homogeneous parallel clustering method is strongly appropriate. The functionality of the framework was based on the efficiency of parallel clustering, and it also notes that operating is the master process and the nodes. Finally, because of the ease of service, the system's assessment is strongly appropriate to consumers in terms of usability.

In addition, the study findings have been established and could be introduced to other universities and schools in the area which will be used as an alternative computer to run application in today's technology demands. This will assist faculty, teachers and staff in researching other technical development and device efficiency.

Nevertheless, with the use of clustering strategies and encouraging e-waste management, universities and schools to alternatively develop outdated computers.
A future collection of machines with changed architectures will be selected for future work to enhance the analysis, in order to observe the effect of heterogeneity on the efficiency and growth of the clustering technique.

Lastly, to increase device reliability, an implementation can require additional measures, configurations, and performance review. Additional testing methods are also recommended to determine the efficiency of the device being built.

References

1. D. Sinha-Khetriwal, The management of electronic waste: a comparative study on India and Switzerland, M.S.
2. M Sang M.Leea & SilvanaTrimi (2016). Innovation for creating a smart future, Journal of Innovation & Knowledge, ISSN: 2444-569X, Vol: 3, Issue: 1, Page: 1-8ay
3. D.D. Williams, C.R. Graham, in International Encyclopedia of Education (Third Edition), 2010
4. Yeo C.S., Buyya R., Pourreza H., Eskicioglu R., Graham P., Sommers F. (2006) Cluster Computing: High-Performance, High-Availability, and High-Throughput Processing on a Network of Computers. In: Zomaya A.Y. (eds) Handbook of Nature-Inspired and Innovative Computing. Springer, Boston, MA. 2006.
5. Mugunthan, S. R. "Novel Cluster Rotating and Routing Strategy for software defined Wireless Sensor Networks." Journal of ISMAC 2, no. 02 (2020): 140-146.

6. Thomas Sterling, in Encyclopedia of Physical Science and Technology (Third Edition), 2003
7. Raj, Jennifer S. "Machine Learning Based Resourceful Clustering With Load Optimization for Wireless Sensor Networks." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 01 (2020): 29-38.
8. Thomas Sterling, ... Maciej Brodowicz, in High Performance Computing, 2018
9. Casey, J. (2015). Computer Hardware: Hardware Components and Internal PC Connections. Guide for undergraduate students. Technological University Dublin
10. Sachidananda Kangovi, in Peering Carrier Ethernet Networks, 2017
11. Naomi J. Alpern, Robert J. Shimonski, in Eleventh Hour Network+, 2010
12. Castaos, I & Garrido, Izaskun & Garrido, Aitor & Sevillano, M. (2009). Design and implementation of an easy-to-use automated system to build Beowulf parallel computing clusters. 1 - 6. 10.1109/ICAT.2009.5348420.
13. Van Steen, M., & Tanenbaum, A. S. (2016). A brief introduction to distributed systems. Computing, 98(10), 967–1009. doi:10.1007/s00607-016-0508-7
14. Matthew L. Massie, Brent N. Chun, and David E. Culler. 2004. The ganglia distributed monitoring system: design, implementation, and experience. Parallel Comput. 30, 7 (2004), 817--840.
15. O'Connor, D.T., and A. Kleyner. 2012. ''Practical Reliability Engineering'', 5th Edition. Chichester, UK: J. Wiley & Sons, Ltd.
16. Petrie, Helen & Bevan, Nigel. (2009). The Evaluation of Accessibility, Usability, and User Experience. C Stepanidis. 10.1201/9781420064995-c20.
17. ISO/IEC 9126, Software Engineering - Product quality, Parts 1-4, 1999-2004.

International Journal of Agricultural Technology 2022Vol. 18(6):2447-2458
Available online http://www.ijat-aatsea.com
ISSN 2630-0192 (Online)

Hatchability of duck eggs as affected by types of incubators under varying relative humidity

Lagasca, A. C.[1*], Beltran, MA. G.[2], Valdez, MT. SJ.[3], Franquera, E. N.[4], Reyno, E. A.[5] and Briones, R. C.[6]

[1]University Extension Center, Central Luzon State University, Science City of Muñoz, Nueva Ecija, Philippines; [2]College of Veterinary Medicine, Tarlac Agricultural University, Camiling, Tarlac, Philippines; [3]Department of Animal Science, College of Agriculture and Forestry, Camiling, Tarlac, Philippines; [4]College of Agriculture and Forestry, Tarlac Agricultural University, Camiling, Tarlac, Philippines; [5]Department of Crop Science, College of Agriculture and Forestry, Tarlac Agricultural University, Camiling, Tarlac, Philippines; [6]Department of Animal Science, College of Agriculture and Forestry, Tarlac Agricultural University, Camiling, Tarlac, Philippines.

Lagasca, A. C., Beltran, MA. G., Valdez, MT. SJ., Franquera, E. N., Reyno, E. A. and Briones, R. C. (2022). Hatchability of duck eggs as affected by types of incubators under varying relative humidity. International Journal of Agricultural Technology 18(6):2447-2458.

Abstract The mean percentage hatchability was not significantly affected by types of incubators, two levels (70% and 80%) of relative humidity, and the interaction effect of types of incubators and relative humidity. However, the percentage of egg hatching was significantly affected by types of incubators such that the means of the percentage hatch in Cabinet-Type Electric Incubator (M=43.95%) and in Bamboo or "Garong"-Type Incubator (M=41.88%) were significantly higher from "Lawanit" Board-Type Incubator (M=27.68%). There was no significant difference in the percentage hatch as affected by two different levels of relative humidity and the interactions of types of incubators and varying percent relative humidity. In this study, the use of Bamboo or "Garong"- Type Incubator indicated the lowest cost (₱0.83) to hatch a duckling, the cheapest (₱17.70) to produce a duckling, and highest ROI of 2.00%.Thus, the Bamboo or "Garong"-Type Incubator was the most economical to use among the three types of incubators.Among the three types of incubators under the two levels of relative humidity, it was observed that the cost to hatch and the cost to produce a duckling is lower under 80% relative humidity with an average cost of ₱1.89 and ₱20.00 respectively. Higher ROI (1.56%) was also observed when the eggs were incubated under 80% RH than 70% RH with an average ROI of 1.54%. Furtheremore, the Cabinet-Type Electric Incubator and Bamboo or "Garong"-Type Incubator were identified to be the most efficient types of incubators. All the hatching parameters were not significantly affected by two levels (70% and 80%) of relative humidity.

Keywords: Bamboo or "Garong" incubator, Duck egg incubation, Parched rice incubation, Rice husk incubation technique

[*] **Corresponding Author:** Lagasca, A. C.; **Email:** armanlagasca@clsu.edu.ph

Introduction

Ducks are commonly raised by rural farmers in the Philippines mostly in Central Luzon and some part of Western Visayas. Around 429,700 families derive their livelihood from it (Santiago, 2018). These farmers, who have less than 100 heads of ducks, contribute to 70% of the ducks in the country (Philippine Statistics Authority [PSA], 2015) as reported by Arrosa (2018). Ducks are preferred by small-holders in the communities compared to other fowls because ducks can adapt and survive under a wide range of climatic conditions, can feed on a variety of feedstuffs, and are resistant to diseases. Raising Mallard Duck (*Anas platyrhynchos*), being known for egg production, provides a good source of income to farmers through its products such as *balut* (embryonated egg) and salted egg. These ducks can also be sold as cull after two years of egg production when their laying performance begins to decline. However, there are some issues and concerns about the duck industry. Results from a farm survey on duck egg production in the Philippines showed that duck farmers generally lack the technical know-how and extension services as well as insufficient supply and high cost of producing good quality ducklings (Chang *et al*., 2008). Most of these problems, particularly on technical and production and supply aspects, are brought about by the issues concerning the type of incubator used and the physical factors to which the eggs were subjected before and during the incubation period.

The duck industry has a promising future given the high demand for salted eggs and *balut* which accounts for 90% of the total egg production in the country (Beltran, 2015). The vast knowledge of local farmers in raising ducks, the availability of complete feeds at different stages of growth (from brooding to laying), and the government programs on strengthening mallard duck production (*Itik Pinas*) all over the country make the duck raising more encouraging. Furthermore, it is one of the special programs of the Bureau of Animal Industry under the Department of Agriculture that aims to contribute to attaining one of the goals of President Duterte on food sufficiency. However, there are other problems with the industry such as the fluctuating prices of eggs, limited space for free-range operations, and inadequate research studies being conducted on duck raising ("Native", 2016).

At present, commercial incubators of varying capacities are being used by *balut* and duck producers in the country. Most of these incubators are operating with electricity for heating and other mechanical functions. However, small-holders experience issues in the cost of procurement, maintenance, and operations. According to Boleli *et al.* (2016), the main focus of research at present is the manipulation of thermal incubation conditions and the integrated

effect of factors that influence incubation. Commercial hatcheries use modern state-of-the-art incubators but one of the questions that need to be answered is how effective and cost-efficient the incubators are in terms of promoting greater hatchability and better chick quality.

Currently, artificial incubators made from locally available materials, using parched rice and rice husk, have been used by hatcheries in Central Luzon and National Capital Region. However, the technology was not widely adopted and little is known about the efficiency of using this type of incubator. Furthermore, literature and studies suggest a wide range of humidity levels inside the incubator and different egg turning frequencies to produce a good hatch. Therefore, it is important to find a method and technique of duck egg incubation that is efficient, less expensive, uses locally available resources, and can easily be adopted by both backyard and commercial raisers.

Thus, this study was conducted to determine the hatchability of duck eggs using different types of incubators under two levels of relative humidity. Also, this study aimed to determine which among the three types of incubators is the most efficient and most economical to be used and be recommended for small-hold, backyard duck raisers, or commercial hatcheries. The results of this study could contribute to the determination of optimum relative humidity which would be needed to attain better hatchability of duck eggs using different types of incubators. Moreover, the results may lead to finding better techniques for hatching duck eggs through the use of locally available resources that would reduce the production cost of quality hatchlings. Through the adoption of technology to be generated, more duck raisers especially the small-hold raisers would be benefited and be able to produce their own ducklings instead of buying them from commercial hatcheries or other suppliers. Furtheremore, the results of this study may be used as a basis for further researches about duck egg incubation.

Materials and methods

Time and place of research

The experiment was conducted at the duck egg hatchery and “balutan” of Mr. Renato C. Ramos in Brgy. Carmen, Zaragosa, Nueva Ecija from June 2020 to July 2020.

Research design

The study was laid out into treatment combinations using a Completely Randomized Design following the 3 x 2 factorial arrangements using three

types of incubators under varying relative humidity. Factor A served as the three types of incubators (Cabinet-Type Electric Incubator, Bamboo or "Garong"-Type Incubator, and "Lawanit" Board-Type Incubator) while the Factor B served as the two levels of relative humidity (70% RH and 80%RH).

Experimental treatments and layout

Five thousand and four hundred duck eggs were used in the study. These were randomly divided into six treatment combinations based on the experimental factor. Each composed of 900 eggs. Each treatment combination was further subdivided into three replications with 300 eggs per replicate.

Setting-up of cabinet-type incubator

Six cabinet-type forced-air electric incubators were used. These were prepared by cleaning especially in the interior area. Thermometers and hygrometers were checked and ensured that these were functioning. The experimental eggs were incubated at a temperature ranging from 37.22-37.78^{o}C.

Setting-up of bamboo or "Garong" and "Lawanit" board incubator

Six "Garong" and six "Lawanit" Board Incubators were also used to complete the types of incubators needed in the study. Each incubator had a diameter of 45-50 cm and a height of 85-90 cm. Nylon net with a size of 75 cm x 70 cm was used to contain 100 eggs and also to contain 1.5 kg of unpolished rice/parched rice as a source of heat.

The rice was heated twice a day to about 42^{0}C to 43^{0}C using a vat or cauldron or "kawa" following the procedures in making *balut* (ATBP.PH, 2016). A pan of water was placed at the bottom of the incubator. Bamboo slats were placed on top of the pan before making a pile of heated rice and duck eggs. Five bags of heated unpolished rice (1.5 kg per bag) and three bags of preheated eggs (100 eggs per bag) were piled in an alternating position having the rice at the bottom and on top when the pile was completed.

The incubators were arranged at a distance of at least four inches from each other. Rice hull was used to fill up spaces between incubators. It served as an insulator and to conserve heat energy.

Egg collection

Large-sized eggs, 65-70 grams, were collected from mated flocks in a selected house on the farm. Eggs laid not later than three days were selected as experimental materials.

Egg setting

The 300 eggs were set in each replication of the three types of incubators. Before setting, eggs were pre-heated under the sun for two to three hours with a temperature ranging from 23.9-26.1^{o}C following the procedures stated in Hatchery Tips (2017).

Incubation duration

Experimental eggs were incubated (with a source of heat) for 15 days. On the 16th day after setting, eggs were transferred onto a table in a closed room (no window) until they were hatched.

Egg turning

Eggs in "Garong" and "Lawanit" board were manually turned two times a day following the procedures in making *balut* by ATBP.PH (2016). On the other hand, eggs in electric incubators were turned by switching on the "turn" button. The eggs were turned twice a day until the 15th day of incubation. From 16 days to hatching, eggs were turned four times within 24 hours.

Candling

The first candling was done on the 10th day after egg setting to determine the number of fertile eggs and infertile eggs. The second candling was done on the 15th day of incubation to select fertile eggs but would fail to hatch due to the following reasons: (1) dead embryo, (2) presence of a red ring or blood around the embryo, (3) enlarged blood vessels, and (4) presence of oozing substance (Smith, 2018).

Relative humidity and temperature control

Relative humidity in the "Garong"-Type and "Lawanit" Board-Type Incubators was controlled by placing a moisture pan inside. Rice was heated two times a day until the 15th day of hatching. After 15 days, rice or palay bags were not heated anymore since the embryos could generate enough heat to keep them warm. However, the humidity in Cabinet-Type Electric Incubator was controlled by placing also moisture pan inside. The temperature was set into 37.5 ^{0}C until 15 days of hatching.

Data analysis

The data from the experiment were subjected to analysis of variance in 3x2 factorial in Completely Randomized Design (Gomez and Gomez, 1984). When significant differences were obtained, means were compared using the Least Significant Difference (LSD) at 5% probability. To facilitate calculations and analysis of experimental data, the computer program Statistical Tool for Agricultural Research was used.

Results

Percentage hatchability

The mean of the percentage hatchability of duck eggs as affected by types of incubators under varying relative humidity is shown in Table 1. The results of the study revealed that percent hatchability of duck eggs was not affected by the types of incubators as indicated by their means (M=48.42%, M=50.40%, and M=37.17%) having no significant difference, $F(2,12) = 2.78$, $p = 0.1020$ when analyzed for variance. The percentage hatchability was not affected also by the two levels (70% and 80%) of relative humidity wherein their means (M=44.98% and M=45.68%) were comparable. As for the effect of the interactions of types of incubators and varying percent relative humidity, it was found out that these interactions did not affect the percent hatchability of duck eggs, $F(2,12) = 1.81$, $p = 0.2057$.

Table 1. Mean of the percentage hatchability of duck eggs as affected by types of incubators under varying relative humidity

Factor A – Types of Incubators	**Factor B – Relative Humidity**		**Factor A Mean**
	70% RH	**80% RH**	
Cabinet-Type Electric Incubator	44.83	52.00	**48.42**
Bamboo or “Garong”-Type Incubator	56.70	44.10	**50.40**
“Lawanit” Board-Type Incubator	33.40	40.93	**37.17**
Factor B Mean	**44.98**	**45.68**	

Percentage hatch

The mean of the percentage hatch of duck eggs as affected by types of incubators under varying relative humidity is presented in Table 2. In this study, the percentage hatch of duck eggs was significantly affected by the types

of incubators, F(2,12) = 5.73, p = 0.0179.The mean percentage hatch in Cabinet-Type Electric Incubators (M=43.95%) and Bamboo or "Garong"-Type Incubators (M=41.88%) was significantly higher than the mean percentage hatch in "Lawanit"Board-Type Incubator (M=27.68%).

However, the percentage hatch was not affected by the two levels (70% and 80%) of relative humidity with which their means (M=38.03% and M=37.65%) were not significantly different, F(1,12) = 0.01, p = 0.9290. As for the effect of the interaction of types of incubators and varying percent relative humidity, it was revealed that these interactions did not affect the percent hatch of duck eggs.

Table 2. Mean of the percentage hatch of duck eggs as affected by types of incubators under varying relative humidity

Factor A – Types of Incubators	Factor B – Relative Humidity		Factor A Mean
	70% RH	80% RH	
Cabinet-Type Electric Incubator	39.00	48.90	**43.95 a**
Bamboo or "Garong"-Type Incubator	48.00	35.77	**41.88 a**
"Lawanit" Board-Type Incubator	27.10	28.27	**27.68 b**
Factor B Mean	**38.03**	**37.65**	

Note: Means followed by the same letter are not significantly different at 5% level of significance by LSD

The most economical type of incubator

The mean of the average cost to hatch a duckling among three types of incubators for 70% and 80% relative humidity is shown in Table 3. In this study, the lowest average cost (₱0.83) to hatch a duckling was determined when using the Bamboo or "Garong"-Type Incubator. It was lower than the cost of using the "Lawanit" Board-Type Incubator and Cabinet-Type Electric Incubator with the average costs of ₱1.12 and ₱4.11, respectively.

Among the three types of incubators under the two levels of relative humidity, it was observed that the lowest average cost (₱0.75) to hatch a duckling was with the use of Bamboo or "Garong"-Type Incubator under 70% while the highest average cost (₱4.47) was by using Cabinet-Type Electric Incubator also under 70% relative humidity. Furthermore, the cost to hatch a ducking was lower (₱1.89) under 80% relative humidity than 70% relative humidity with an average of ₱2.59.

Table 3. Mean average cost (₱) to hatch a duckling using three types of incubators under 70% and 80% relative humidity

Incubator type	Cost to hatch a duckling (₱)		Mean
	70% RH	80% RH	
Cabinet-Type Electric Incubator	4.47	3.75	**4.11**
Bamboo or "Garong"-Type Incubator	0.75	0.91	**0.83**
"Lawanit" Board-Type Incubator	1.22	1.01	**1.12**
Mean	**2.15**	**1.89**	

The mean of the average cost to produce a duckling among three types of incubators for 70% and 80% relative humidity is shown in Table 4. The study revealed that the Bamboo or "Garong"-Type Incubator was the cheapest to use among the three types of incubators to produce a duckling with an average cost of ₱17.70. However, a higher average cost to produce a duckling was realized with the use ofCabinet-Type Electric Incubator and "Lawanit" Board-Type Incubator with the average costs of ₱21.04 and ₱23.72, respectively.

Among the three types of incubators under the two levels of relative humidity, it was observed that the lowest average cost (₱16.02) to produce a duckling was by using Bamboo or "Garong"-Type Incubator under 70% while the highest average cost (₱26.02) was by using the "Lawanit" Board-Type Incubator also under 70% relative humidity. Moreover, it was determined that it was cheaper ((₱20.00) to produce a duckling under 80% relative humidity than under 70% relative humidity with an average cost of ₱21.64.

Table 4. Mean of the average cost (₱) to produce a duckling among three types of incubators for 70% and 80% relative humidity

Incubator type	Cost to hatch a duckling (₱)		Mean
	70% RH	80% RH	
Cabinet-Type Electric Incubator	22.87	19.21	**21.04**
Bamboo or "Garong"-Type Incubator	16.02	19.37	**17.70**
"Lawanit" Board-Type Incubator	26.02	21.42	**23.72**
Mean	**21.64**	**20.00**	

The mean of the average percentage ROI among the three types of incubators for 70% and 80% relative humidity is illustrated in Table 5. Among the three types of incubators, the Bamboo or "Garong"-Type Incubator obtained the highest average percentage return on investment (2.00%) while the "Lawanit" Board-Type Incubator obtained the lowest average ROI (1.21%).

Among the three types of incubators under the two levels of relative humidity, the highest average percentage ROI (2.37%) was obtained when Bamboo or "Garong"-Type Incubator was used under 70%. However, the

lowest average percentage ROI (0.99%) was attained when "Lawanit" Board-Type Incubator was used under 70% relative humidity. On the other hand, higher ROI (1.56%) was realized under 80% relative humidity than under 70% relative humidity with an average ROI of 1.54%.

Table 5. Mean of the average percentage (%) return on investment among three types of incubators for 70% and 80% Relative Humidity

Incubator type	% ROI		Mean
	70% RH	80% RH	
Cabinet-Type Electric Incubator	1.25	1.62	**1.44**
Bamboo or "Garong"-Type Incubator	2.37	1.62	**2.00**
"Lawanit" Board-Type Incubator	0.99	1.43	**1.21**
Mean	**1.54**	**1.56**	

The most efficient type of incubator

Based on the results of the study, only the percentage hatch was affected by a certain factor - the types of the incubators. The analysis of variance yielded a main effect for the type of incubator, $F(2,12) = 5.73$, $p < .05$, such that the average percentage hatch was significantly higher inCabinet-Type Electric Incubator (M=43.95%) and Bamboo or "Garong"-Type Incubator (M=41.88%) than in"Lawanit"Board-Type Incubator (M=27.68%) (see Table 2). The main effects of humidity and interaction were non-significant, $F(1,12) = 0.01$, $p > .05$ and $F(2,12) = 2.27$, $p > .05$, respectively. Therefore, the most efficient types of incubators are the Cabinet-Type Electric Incubator and the Bamboo or "Garong"-Type Incubator.

Discussion

Commercial incubators of varying capacities are being used by *balut* and duck producers. However, the cost of procurement and operation, and the effectiveness and cost-efficiency in terms of promoting greater hatchability of these incubators are the major concerns needed to be answered. Artificial incubators made from locally available materials are being used by hatcheries in Central Luzon and National Capital Region. However, the technology was not widely adopted and little is known about the efficiency of using this type of incubator. The study was conducted to determine the effect of types of incubators under varying relative humidity on the different hatching parameters.

This study demonstrated that the hatchability of fertile duck eggs was not affected by types of incubators ($p = 0.1020$), by the two levels (70% and 80%)

of relative humidity, and the interactions of types of incubators and varying percent relative humidity (p = 0.2057) when analyzed for variance.This result is parallel with the study of Indarsih *et al.* (2019) which revealed that the sawdust incubator gave similar fertility, hatchability, and embryonic mortality values as the electric incubator. Also, this result is associated with the findings of the study Bruzual *et al.* (2000), whichpointed out that fertile hatchability was optimum when incubated at 53% relative humidity. The findings of Bruzual *et al.* (2000), is supported by Hitchener (2017) and Daniels (2020) who recommended that the ideal relative humidity is at 55%.

The percentage hatch of duck eggs was significantly affected by the types of incubators (p = 0.0179). In this experiment, the mean percentage hatch in Cabinet-Type Electric incubators (M=43.95%) and Bamboo or "Garong"-Type Incubators (M=41.88%) was significantly higher than the mean percentage hatch in "Lawanit" Board-Type Incubator (M=27.68%). Boleli *et al*. (2016) explained that this is because the latter provides better incubation physical conditions such as ventilation, egg turning, and egg position, which may affect hatchability. However, the percentage hatch was not affected by the two levels (70% and 80%) of relative humidity and the interaction of types of incubators and varying percent relative humidity. These results could be correlated to the statement of Paniago (2005) as specified by Boleli *et al.* (2016) that despite the technological advances of modern incubation machines, still, the quality of labor both inside and outside the hatcheries determines the success of incubation.

The most economical type of incubator was determined based on the following aspects: (1) cost to hatch a duckling; (2) cost to produce a duckling; and (3) percentage ROI.In this study, the use of Bamboo or "Garong"-Type Incubator indicated the lowest cost (₱0.83) to hatch a duckling, the cheapest (₱17.70) to produce a duckling, and has the highest ROI of 2.00%.

Among the three types of incubators under the two levels of relative humidity, it was observed that the cost to hatch and the cost to produce a duckling was lower under 80% relative humiditywith an average cost of ₱1.89 and ₱20.00 respectively against ₱2.15 and ₱21.64 under 70% relative humidity. Higher ROI (1.56%) was also observed when the eggs were incubated under 80% RH than 70% RH with an average ROI of 1.54%. Moreover, the determination of most economical type of incubator is greatly affected by the hatchability of fertile duck eggs from the specific incubator. This means that the higher the hatchability of an egg from a certain incubator, the lower the cost that may incur to hatch or to produce a duckling. These results are supported by the results of the experiments conducted by El-Hanoun and Mossad (2008) pointing out that the hatchability of fertile Pekin Duck eggs could be improved

by raising the relative humidity (RH) to 80% during the period of 14-28 days of incubation. Their experiments are related to the study of Onbasilar *et al*. (2014) which revealed that hatchability of set and fertile eggs of Pekin Ducks were higher when incubated at 37.5^0C and sprayed with warm water ($25\text{-}28^0$C) from day 4 to day 25 of incubation.

The most efficient type of incubator was determined only when the types of incubators, relative humidity, and the interaction effect of incubator and humidity have a significant effect on the hatching parameters. These conditions were discussed by Boleli *et al.* (2016) in their article regarding optimizing production efficiency that includes manipulation of thermal incubation conditions and the integrated effect of factors that influence incubation. In this present study, only the percentage hatch was affected by a certain factor, the types of the incubators, which revealed that a significantly higher percentage hatch was obtained in Cabinet-Type Electric Incubator and Bamboo or "Garong"-Type Incubator than in "Lawanit" Board-Type Incubator. In addition, all the hatching parameters were not significantly affected by two levels of relative humidity. Therefore, the effects of 70% and 80% relative humiditiesare comparative.

In summary, the study showed that the Cabinet-Type Electric Incubator and Bamboo or "Garong"-Type Incubator yield a significantly higher number of hatch eggs than "Lawanit" Board-Type Incubator. Also, the Cabinet-Type Electric Incubator and Bamboo or "Garong"-Type Incubators were the most efficient types of incubators. Bamboo or "Garong"-Type Incubator was the most economical (lowest cost to hatch and to produce duckling and highest % ROI) type of incubator to use.

Acknowledgements

The author expresses his deepest appreciation and sincerest thanks to Dr. Ma. Asuncion G. Beltran, his adviser, for her support, suggestions, and advice towards the fulfilment of his study; Mr. Renato C. Ramos, owner of the duck egg hatchery where the study was conducted, for his full support in the study, efforts in preparation of the site, and provision of hatching materials and equipment; Ms. Elisa E. Mallari, Regional Livestock Coordinator of the Department of Agriculture-Regional Filed Office III, for her invaluable support and providing duck egg incubators; and his CLSU supervisor, Dr. Eugenia G. Baltazar, Director for Extension of Central Luzon State University, for her unselfish support by providing him minimal workload to finish the study.

References

Arrosa, M. A. S. and Piadoza, M. E. S. (2018). Analysis of marketing options of duck egg producers in Laguna. Philippine Journal of Veterinary and Animal Sciences, 44:111-121.

ATBP. PH. (2016). How to make balut egg (Pinoy balut/ Filipino balut). Retrieved from https://www.atbp.ph/2016/07/11/make-balut-egg-pinoy-balut-filipino-balut/

Beltran, MG. (2015). Advances in duck raising. *Lectures in Advances in Poultry Production and Management.* TAU, Camiling, Tarlac, Philippines

Boleli, I. C., Morita, V. S., Matos Jr., J. B., Thimotheo, M. and Almeida, V. R. (2016). Poultry egg incubation: Integrating and optimizing production efficiency. Brazilian Journal of Poultry Science. Retrieved from http://dx.doi.org/10.1590/1806-9061.

Bruzual, J. J., Peak, S. D. and Peebles, E. D. (2000). Effects of relative humidity during incubation on hatchability and body weight of broiler chicks from young breeder flocks. Poultry Science, DOI:10.1093/ps/79.6.827. Retrieved from https://pubmed.ncbi.nlm.nih.gov/10875763/

Chang, H. S., Villano, R., Velasco, ML., De Castro, N. L. and Lambio, A. L. (2008). Duck egg production in the Philippines: Results from a farm survey. Philipp. Journal of Veterinary Medicine, 45:95-108.

Daniels, T. (2020). Incubation humidity. *Incubation Humidity Guide.* Retrieved from https://poultrykeeper.com/incubation-brooding/incubation-humidity/

El-Hanoun, A. M. and Mossad, N. A. (2008). Hatchability improvement of peking duck eggs by controlling water evaporation rate from egg shell. Egypt Poultry Science, 28767-784.

Gomez, K. A. and Gomez, A. A. (1984). Statistical procedures for agricultural research. Retrieved from https://books.google.com.ph/books

Hatchery tips. (2017). Retrieved from www.hatchability.com/Aviagen-hatchery.pdf

Hitchener, G. (2017). Hatching duck eggs. *Cornell University College of Veterinary Medicine.* Retrieved from https://www.vet.cornell.edu/animal-health-diagnostic-center/programs/duck-research-lab/hatching-duck-eggs

Indarsih, B., Sukartha Jaya I. N. and Mutmainah, A. (2019). Sawdust incubator: An alternative traditional hatchery technique for Japanese quails (Coturnix japonica). Livestock Research for Rural Development, 32. Retrieved from Irrd.cipav.org.co

Native duck raising. (2016). Retrieved from https://www.pressreader.com

Onbasilar, E. E., Erdem, E., Kocakaya, A. and Hacan, O. (2014). Effect of spraying Pekin duck eggs obtained from different breeder age on hatchability. European Poultry Science (EPS). Retrieved from https://www.European-poulrty-science.com.

Paniago, M. (2005). Artificial incubation of poultry eggs-3,000 years of history. Semantic Scholar. Retrieved from https://www.semanticscholar.org/paper

Santiago, R. C. (2018). Duck egg business: "Paunlarinsatamangpagpapalahi, pagkain, at pag-aalaga." Lecture in Duck Egg Production.ppt.

Smith, K. (2018). How to spot a bad egg in the incubator. Backyard Chicken Coops. Retrieved from https://www.backyardchickencoops.com.au

(Received: 5 May 2022, accepted: 12 October 2022)

Webserver Utilization Using Common of-the-shelf Computers

Jeffrey John R. Yasay
Tarlac Agricultural University
Camiling Tarlac, Philippines
88088088088088@yahoo.com

ABSTRACT

The study focused on the collection of condemned computers that would be linked to form clustered computers. This study is limited in the utilization of a web server using common off-the-shelf computers which was conceptualized to provide not only a web server but also other services (e.g. opening multiple applications, web browsing, web site accessing) in a more cost effective manner. The utilization of web server using common off-the-shelf computers was based on the cluster framework. The effectiveness of parallel computing on the web server as evaluated by the IT experts was confirmed. The development of the system was based on rapid application development model. The utilization of web server used old homogeneous and diskless set of machines. The system performance was not degraded during the duration of the evaluation and monitoring. Achieving tremendous machine power with reduced development costs was confirmed. The system is highly acceptable based on experts' opinion. The functionality of the system was based on rates, latency and scalability. Security design was evaluated as excellent based on firewall, user authentication, limitations of access to files and confidentiality of records. The system is highly acceptable to the users as to accessibility because of the simplicity of operations and flexibility of the system such as administering multiple applications, opening and accessing the university website.

Keywords

Web server, homogenous, diskless, off-the-shelf, parallel computing.

1. INTRODUCTION

Before, computers were just regarded as one form of luxury, minds have been set that only those who belong in the higher level economic strata can actually take hold or own computers. As technology continues its powerful outgrow, millions of people easily adapt it to suffice the requisition of every demand.

Technology is constantly evolving and growing, and it is inevitable that this progression will continually offer new and interesting advances in our world. [1]. Computers that are purchased then are inevitably outmoded in terms of speed and their ability to process new applications. When this happens, the old computers are to be deemed obsolete. This also holds true in businesses where computing is playing key roles in their growth and success. As necessity dictates, there is a need to find a way in which these computers, deemed obsolete and useless, will be useful to build computers capable of meeting the demands of whatever endeavors. Cluster is a type of parallel or distributed computer system, which consists of a collection of inter-connected stand-alone computers working together as a single integrated computing resource [2], [3]. A common use of cluster computing is to balance load traffic on high-traffic websites.

Internet usage has manifold, serving humongous amount of content every day. The major contribution to this exploding user base has been driven from emerging markets, especially from Asia, currently contributing over 50% [4]. The proportion of Internet users has steadily increased to more than 90% in many economically developed countries [15]. However, as demand and traffic increases, more and more sites are challenged to keep up, literally, particularly during peak periods of activity. Downtime or even delays can be disastrous, forcing customers and profits to go elsewhere. The solution then, to the budget-constraint organization, is to develop a more scalable web server using common-off-the-shelf computers. A web server is the combination of computer and the program installed on it. Web server interacts with the client through a web browser. It delivers the web pages to the client and to an application by using the web browser and the hypertext transfer protocol (HTTP) respectively. [5].

Internet is the easiest way to find information about any kind of organization, and the first impression about an organization is almost always based on its Web site. [6]. But time spent preparing servers, publishing codes, dealing with deployment issues- as opposed to throwing money at unreliable oversold servers are some of the hindrances of achieving highly available and highly scalable web services.

With this, the utilization of web server using common off-the-shelf computers was conceptualized to provide a more scalable web server at a minimal cost. Also, this study was conducted to utilize other functions of web server.

2. THE DESIGN AND ARCHITECHTURE

Setting-up the clustered web sever

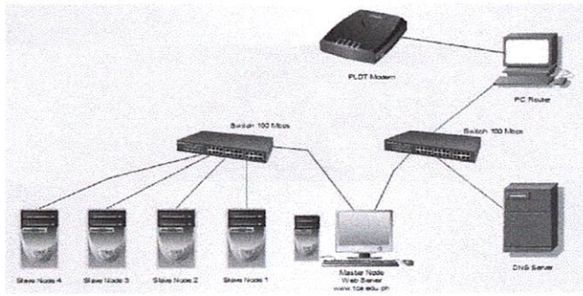

Figure 1. Connectivity of Cluster

Figure 1 shows the connectivity of clustered web server using common of-the-shelf computers.

The approach utilized for the study is the research and development method. The approach was chosen due to proponent's consciousness of a need to design a Web Server Using Common Off-The-Shelf Computers. The system was based on the parallel computing framework. The Clusters work customarily on master-slave where master has the duty to receive jobs and distribute them into subtasks to run on slaves. All slaves return their work back to

master, master may host multiple users and each user in return may submit different processed jobs to master.

2.1 Tools Used for the Utilization of Web Server Using Common Off-the-Shelf Computers

Hardware and Software requirement specifications were used as basis in designing the system. Utilization of Web Server Using Common Off-The-Shelf Computers was developed using the following software and hardware:

Personal computer (PC) - a digital computer designed for use by only one person at a time. A typical personal computer assemblage consists of a central processing unit (CPU), which contains the computer's arithmetic, logic, and control circuitry on an integrated circuit; two types of computer memory, main memory, such as digital random-access memory (RAM), and auxiliary memory, such as magnetic hard disks and special optical compact discs, or read-only memory (ROM) discs (CD-ROMs and DVD-ROMs); and various input/output devices, including a display screen, keyboard and mouse, modem, and printer.[7].

Fast Ethernet Switch - provided 10 times higher bandwidth, and other new features such as full-duplex operation, and auto-negotiation. [11].

Straight Through - a straight through Ethernet cable is used and the pairs will match up properly. [12].

Ubuntu ABC GNU/Linux - based distribution allows to automatically build Beowulf clusters either live or installing the software in the frontend.[8].

Apache Web Server - a free open source code that has been ported over to many platforms such as Linux and allows anyone to make modification to the server.[14].

MySQL - A relational database management system (RDBMS). MySQL stands for "My Structured Query Language". A database enables you to efficiently store, search, sort, and retrieve data.[13].

PHP - A server-side scripting language designed specifically for the Web.[13].

2.2 Hardware Set Up Procedures

To maximize the benefits of the system, homogenous computing is being used to interconnect similar processing cores or units to build a high performance computer.[10]. All nodes and the master node are all in the same specifications.

2.3. Software Installation

After the machines were assembled, the next step is to install the software. There are many distribution of Linux available and different people prefer different distributions for different reasons. The proponent used ABC GNU Linux (Ubuntu 9.04) [8] with the default kernel as a basis.

Before installing ABC GNU Linux (Ubuntu 9.04), the proponent gathered information about the hardware specifications. The following are the detailed information about the hardware used for the system.

2.4. System Specifications and Testing

2.4.1 Master Node (Node 1)

Processor: Intel Celeron M

CPU Number: 1

CPU Speed: 2266 MHz

Total Memory: 509452 Kb

Hard Disk Size: 40 Gb

Built-in LAN Card: Intel UNDI, PXE-2.0 (build 082)

Copyright © 1997, 1998, 1999 Intel Corporation VIA Rhine II Fast Ethernet Adapter V2.38 (2004/09/15)

MAC Address: 00:48:54:64:E3:F5

External LAN Card: SiS (Model 2003)

MAC Address: 00:13:D3:CF:AD:24

parts

Slave Nodes	Specifications	Details/Amount
1	Memory	190660Kb
	MAC Address	00:13:03:CF:AE:C6
2	Memory	190660Kb
	MAC Address	00:13:D3:CF:AA:F7
3	Memory	255684Kb
	MAC Address	00:13:D3:CF:AD:5F
4	Memory	255684Kb
	MAC Address	00:13:D3:CF:AF:77

Table 1. Specification and details of slave nodes.

Table 1 shows the specification and details of slave nodes. The rest of the master node specification and details are the same.

2.5 Cluster Configuration

2.5.1 Installation of ABC GNU Linux Kernel ISOLINUX3.63 Debian Master Node

1. Boot from the CD-ROM
2. Choose an install mode, press enter then follow the directions on the screen.
3. Select use entire disk to partition the hard disk.
4. Create username and password
5. Install ABC GNU (Ubuntu 9.04)
6. Installation of Apache 2.0

```
sudo apt-get install apache2
sudo /etc/init.d/apache2 start
sudo /etc/init.d/apache2 stop
```

7. Installation of PHP 5

```
sudo apt-get install php5 libapache2-mod-php5
sudo /etc/init.d/apache2 restart
```

8. Installation of mysql

```
sudo apt-get install mysql-server
```

mysql> SET PASSWORD FOR 'root'@'localhost' =PASSWORD('xxxxxx')

2.5.2 Configuration of the nodes CMOS

Slave 1 (Node 2)
Configuration of CMOS
- Halt on ALL ERROR
- Set-up to Boot from Network

Slave 2 (Node 3)
Configuration of CMOS
- Halt on ALL ERROR
- Set-up to Boot from Network

Slave 3 (Node 4)
Configuration of CMOS
- Halt on ALL ERROR
- Set-up to Boot from Network

Slave 4 (Node 5)
Configuration of CMOS
- Halt on ALL ERROR
- Set-up to Boot from Network

2.6 Cluster Loading

This is the command and syntax in the cluster process done on command line interface (CLI)

2.6.1 Master Node

master@master-desktop:~$ ifconfig

eth0 Link encap:Ethernet HWaddr 00:48:54:64:e3:f5

inet addr:192.168.0.1
Bcast:192.168.0.255 Mask:255.255.255.0

inet6 addr: fe80::248:54ff:fe64:e3f5/64 Scope:Link

UP BROADCAST RUNNING MULTICAST MTU:1500 Metric:1
RX packets:0 errors:0 dropped:0 overruns:0 frame:0
TX packets:132 errors:0 dropped:0 overruns:0 carrier:0
collisions:0 txqueuelen:1000
RX bytes:0 (0.0 B) TX bytes:10187 (10.1 KB)
Interrupt:18 Base address:0xe000

eth1 Link encap:Ethernet HWaddr 00:13:d3:cf:ad:24
inet addr:192.168.100.133
Bcast:192.168.100.255 Mask:255.255.255.0
inet6 addr: fe80::213:d3ff:fecf:ad24/64 Scope:Link

UP BROADCAST RUNNING MULTICAST MTU:1500 Metric:1
RX packets:776 errors:0 dropped:0 overruns:0 frame:0
TX packets:162 errors:0 dropped:0 overruns:0 carrier:0

collisions:0 txqueuelen:1000

RX bytes:105830 (105.8 KB) TX bytes:15957 (15.9 KB)
Interrupt:23

lo Link encap:Local Loopback
inet addr:127.0.0.1 Mask:255.0.0.0
inet6 addr: ::1/128 Scope:Host
UP LOOPBACK RUNNING MTU:16436 Metric:1
RX packets:494 errors:0 dropped:0 overruns:0 frame:0
TX packets:494 errors:0 dropped:0 overruns:0 carrier:0
collisions:0 txqueuelen:0
RX bytes:128398 (128.3 KB) TX
bytes:128398 (128.3 KB)
master@master-desktop:~$

2.6.2 Test Master (Node 1)

master@master-desktop:~$ cat clusterhosts
192.168.0.1
master@master-desktop:~$

2.6.3 Test the cluster system

(Node1, Node 2, Node 3, Node 4 and Node 5)

master@master-desktop:~$ cat clusterhosts
192.168.0.1
192.168.0.13
192.168.0.3
192.168.0.10
192.168.0.8
master@master-desktop:~$

2.6.4 Test the Base linear of the Master and Slave Nodes

master@master-desktop:~$ recon -v clusterhosts
n-1<4633> ssi:boot:base:linear: booting n0 (192.168.0.1)
n-1<4633> ssi:boot:base:linear: booting n1 (192.168.0.13)
n-1<4633> ssi:boot:base:linear: booting n2 (192.168.0.3)
n-1<4633> ssi:boot:base:linear: booting n3 (192.168.0.10)
n-1<4633> ssi:boot:base:linear: booting n4 (192.168.0.8)
n-1<4633> ssi:boot:base:linear: finished

Woo hoo!

recon has completed successfully. This means that you will most likely be able to boot LAM successfully with the "lamboot" command (but this is not a guarantee). See the lamboot(1) manual page for more information on the lamboot command.

If you have problems booting LAM (with lamboot) even though recon worked successfully, enable the "-d" option to lamboot to examine each step of lamboot and see what fails. Most situations where recon succeeds and lamboot fails have to do with the hboot(1) command (that lamboot invokes on each host in the hostfile).

---

master@master-desktop:~$

2.6.5 Test the LAM Boot

master@master-desktop:~$ lamboot -v clusterhosts

LAM 7.1.2/MPI 2 C++/ROMIO - Indiana UniverResity

n-1<4651> ssi:boot:base:linear: booting n0 (192.168.0.1)

n-1<4651> ssi:boot:base:linear: booting n1 (192.168.0.13)
n-1<4651> ssi:boot:base:linear: booting n2 (192.168.0.3)
n-1<4651> ssi:boot:base:linear: booting n3 (192.168.0.10)
n-1<4651> ssi:boot:base:linear: booting n4 (192.168.0.8)
n-1<4651> ssi:boot:base:linear: finished

master@master-desktop:~$

2.6.6 Cluster view of GANGLIA monitoring tool (GUI)

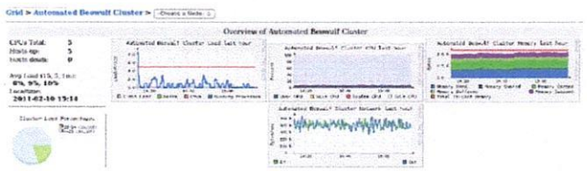

Figure 3. Cluster view using Ganglia

The figure 3 shows the cluster view of the system. At the top of the figure, are the CPU counts and the summary graphs for the cluster. The pie chart shows how much of the cluster is busy. A set of small graphs show the node processing of each CPU.

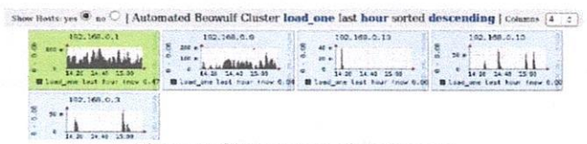

Figure 4. Cluster view of each nodes

The figure 4 shows the cluster view of each nodes. Master node has the ip address of 192.168.0.1 with 4 slave nodes. It also shows the different reading process of each slave nodes.

2.6.7 Website access from external

Figure 5. Tarlac Agricultural University website

The figure 5 shows the running web site which is being access thru world wide web (www.tau.edu.ph).

3. Results of the evaluation

Evaluation of the utilized web server consisted of two processes: by IT experts and by the users. The IT experts assessed the system as to: functionality, reliability and security while the users rated the system as to accessibility. The Ganglia monitoring tool [9] was used as the basis in evaluating and monitoring phase, snapshots of the system were taken every hour and saved to monitor files. In addition, snapshots of the website running on the Master Node (node 1) were taken every hour to test the functionality of the system. Also to test the flexibility of the system, snapshots on the administering of multiple applications were taken.

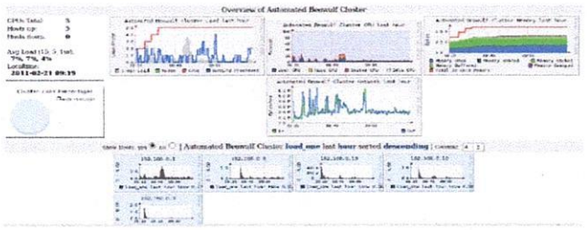

Figure 6. Snapshot of running Ganglia monitoring tool

Figure 6 shows a screen shot in evaluating the effectiveness of parallel computing on the system, physical view of the clustered system were taken to monitor the network packets, disk write and memory usage of the system.

On the eight week (8) of monitoring, the hard disk of the Master Node (Node 1) and the memory of Slave 4 (Node 5) failed due to power failure. Also, the system failure occurred because of the environmental constraints of the location where the study was conducted. The temperature, radiation exposure and corrosive exposure are often critical for computer-based systems. Furthermore, the study used old computers, because of component interdependencies like the hard disk and memory usage, faults could happen anytime.

3.1 Evaluation of the System by IT Experts

Six Information Technology experts of renowned expertise in systems development judged the system. Their comments and suggestions were considered in the improvement of the system.

Evaluation Criteria	Mean	Descriptive Rating
Functionality of the System	4.44	Very Good
Security of the System	4.83	Excellent
Reliability of the System	4.44	Very Good

Table 2. Evaluation Criteria for IT Experts

Table 2 shows the criteria of IT experts. Functionality of the system has a composite mean of 4.44, security of the system

has a composite mean of 4.83 and reliability of the system has a composite mean of 4.44. It proves that the system passed the criteria coming from the IT experts.

3.2 Evaluation of the System by Users

Evaluation Criteria	Mean	Descriptive Rating
System Accessibility	4.69	Excellent
Reliability of the System	4.08	Very Good

Table 3. Evaluation Criteria for Users

Table 3 shows the evaluation result of the system by the users. It shows that system accessibility has a

5. Conclusion and Future Work

Based on the findings of the study, it was concluded that the web server used an old homogeneous (the master and all slaves are of same brand model) and diskless set of machines for the experimental set-up. The system performance was not degraded during the duration of the evaluation and monitoring. Achieving functional machines with reduced development costs was confirmed. Moreover, the utilization of web server using common off-the-shelf computers is highly acceptable based on experts' opinion. The functionality of the system was based on rates, latency and scalability. Security design was evaluated as excellent based on firewall, user authentication, limitations of access to files and confidentiality of records. Lastly, the web server utilization using common off-the-shelf computers is highly acceptable to the users as to accessibility because of the simplicity of operations and flexibility of the system such as administering multiple applications, opening and accessing the university website.

For future work to improve the study, future set of machines having different configurations shall be chosen to observe the effect of heterogeneity on the performance of the clustered system. Furthermore, implementation may require additional steps, configurations and performance analysis to improve the system stability. It is as well, recommended to find other monitoring tools to evaluate the performance of the system.

5. Acknowledgment

This study would not be possible without the support of the Tarlac Agricultural University Tarlac, Philippines.

6. References

[1] Jones, V. & Jo, J.H. (2004). Ubiquitous learning environment: An adaptive teaching system using ubiquitous technology. In R. Atkinson, C. McBeath, D. Jonas-Dwyer & R. Phillips (Eds), Beyond the comfort zone: Proceedings of the 21st ASCILITE Conference (pp. 468-474. Perth, 5-8 December. http://www.ascilite.org.au/conferences/perth04/procs/jones.html.

[2] G.F. Pfister, In Search of Clusters, 2nd Edition, Prentice Hall PTR, pp 29, 1998, ISBN 0-13-899709-8.

[3] R. Buyya (editor), High Performance Cluster Computing: Architectures and Systems, Vol. 1, Prentice Hall PTR, NJ, USA, 1999.

[4] World internet usage and population statistics, [online] Available: http://www.internetworldstats.com.

[5] Rashmi Kanta. Das. 2014. J2EE made easy, New Delhi: Vikas Publishing House PVT LTD.

[6] Ismailova, R. & Kimsanova, G. Univ Access Inf Soc (2017) 16: 1017.

[7] The Editors of Encyclopaedia Britannica. 2017. Personal computer. (December 2017). Retrieved June 8, 2018 from https://www.britannica.com/technology/personal-computer

[8] Castaos, I & Garrido, Izaskun & Garrido, Aitor & Sevillano, M. (2009). Design and implementation of an easy-to-use automated system to build Beowulf parallel computing clusters. 1 - 6. 10.1109/ICAT.2009.5348420.

[9] Matt Massie. 2013. Monitoring with Ganglia, Sebastopol, CA: OReilly.

[10] Y. Gao and P. Zhang, "A Survey of Homogeneous and Heterogeneous System Architectures in High Performance Computing," 2016 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, 2016, pp. 170-175.

[11] Irimia R, Gottschling M (2016) Taxonomic revision of Rochefortia Sw. (Ehretiaceae, Boraginales). Biodiversity Data Journal 4: e7720.

[12] W. Wu, "The Design of a New Network Cabling Experimental Instrument Based on Embedded System." Advanced Materials Research 328, pages 2427-2431, 2011

[13] Luke Welling and Laura Thomson. 2003. PHP and MySQL Web Development, Second Edition, Sams.

[14] Fatma Mbarek, Volodymyr Mosorov, and Rafał Wojciechowski. 2017. Web Server Latency Reduction Study. Informatics Control Measurement in Economy and Environment Protection 7, 2 (2017), 19–23. DOI:http://dx.doi.org/10.5604/01.3001.0010.4829

[15] Mohammad Salahuddin and Khorshed Alam. 2016. Information and Communication Technology, electricity consumption and economic growth in OECD countries: A panel data analysis. International Journal of Electrical Power & Energy Systems 76 (2016), 185–193. DOI:http://dx.doi.org/10.1016/j.ijepes.2015.11.005

Survey of Physical, Chemical and Microbial Water Quality of Irrigation Sources in Tarlac, Philippines

Edmar N. Franquera, Cielito A. Beltran, Ma. Asuncion G. Beltran and Ruth Thesa B. Franquera

Abstract The main sources of irrigation water for irrigating crops comes from major rivers. Usually these water sources which can be used for irrigating various crops could be very vulnerable to contamination. The aim of the study was to determine the physical, chemical and microbial water quality of the different irrigation sources in Tarlac and to compare it with the existing water quality guidelines stipulated in the DENR AO 08 Series of 2016. The water samples collected from the surface water of different rivers were subjected to laboratory analysis. Higher TSS was found to be during wet season as compared during the dry season. Higher COD was found both in dry and wet seasons in Benig river. All of the major rivers have a less than 0.05 mg/l lead and 0.0002 mg/l mercury based from the result of the laboratory analysis. The highest dissolved oxygen was found to be within the Tarlac River both during the dry and wet season. Comparing with the National standards from the DENR the major rivers of Tarlac surpasses the minimum standards of classification of water bodies with dissolved oxygen ranging from 2 to 6 mg/l. The lowest dissolved oxygen was found in Concepcion River during the dry season (5.0 mg/l) and in Rio Chico River (4.8 mg/l) during the wet season. Higher total dissolved solids were observed in the different rivers during the dry season which ranges from 300 to 560 mg/l as compared during the wet season which ranges from 169 to 540 mg/l respectively. The nitrate concentrations of the different rivers in Tarlac shows to be within the range of the National Standards of the DENR. Higher concentrations of *E. coli* and fecal coliform count were also noted within the different rivers of Tarlac.

Keywords Water quality · River · Irrigation · Tarlac

E. N. Franquera (✉) · C. A. Beltran (✉) · Ma. A. G. Beltran (✉) · R. T. B. Franquera (✉)
Tarlac Agricultural University, Malacampa, Camiling, Tarlac, Philippines
e-mail: edmarfranquera123@gmail.com

C. A. Beltran
e-mail: tolitsbeltran@yahoo.com

Ma. A. G. Beltran
e-mail: marizonbeltran@yahoo.com

R. T. B. Franquera
e-mail: edmarfranquera@yahoo.com

© Springer Nature Switzerland AG 2019

R. Sun and L. Fei (eds.), *Sustainable Development of Water and Environment*, Environmental Science and Engineering,
https://doi.org/10.1007/978-3-030-16729-5_2

edmarfranquera123@gmail.com

1 Introduction

Water is life. All living organisms on earth need fresh water. The major user of freshwater in most countries is agriculture. The largest single user of freshwater in the world today which consumes an average of 70% globally is accounted in agriculture.[1] However, the availability of freshwater is already decreasing due to water pollution. Agriculture is considered to be a casualty of water pollution but it also causes and contributes to water pollution due to excess nutrients by too much application of fertilizers, excessive use of pesticides and other pollutants. Globally, agriculture is also considered to be the major cause of degradation of surface including groundwater resources as a result of erosion, excessive farming contaminating freshwater like wastewater coming from large poultries and piggeries, chemical run off and other indiscriminate human activities and improper agricultural management practices. Waste coming from swine is significant source of fecal pollution leading to water pollution by contaminating of ground and surface water from lagoon overflow and the use of lagoon surface water for irrigation. Thus, it is important to test a system or test a technology such as potential aquatic plants to decontaminate the wastewaters so that this will resolve the problem.

In the Philippines, agriculture wastewater is one of the major sources of water pollution which accounted 37%.[2] In addition, only 10% of wastewater is treated while 58% of groundwater is contaminated. Regions which had unsatisfactory ratings for their water quality criteria include National Capital Region (NCR), Southern Tagalog Region, Central Luzon (Region 3) and Central Visayas. Hence, there is a need to address the global implications of water quality and there is a need for wastewater treatments. In central Luzon, the agricultural land area is 653,607 km^2 and 9.1% contributed to the agricultural BOD generation, 9.0% industrial BOD generation and 9.9% domestic BOD generation leading to water quality degradation and contamination.[3]

Generally, the availability of clean freshwater is becoming a primary limitation to human activities expansion and also the scope or capacity of our agricultural lands to feed the tremendous population growth not only in the Philippines but globally. There are an estimated 2.2 million metric tons of organic water pollution that occur in the Philippines each year and the annual economic losses caused by water pollution are estimated at Php67 Billion which is equivalent to more or less US$1.3 billion.[4] Hence, this study aims to quantify the physical, chemical and microbiological water qualities of the different river waters in Tarlac, Philippines.

[1] www.fao.org. Last accessed 30 Nov 2017.

[2] www.greenpeace.org. Last accessed 30 Nov 2017.

[3] www.wipo.int/wipo_ip_mnl_15_t4. Last accessed 27 Nov 2017.

[4] www.wepa-db.net.philippines.overview. Last accessed 30 Nov 2017.

edmarfranquera123@gmail.com

2 Methodology

2.1 *Gathering/Collection of Data of Existing Irrigation Water Sources in Tarlac*

The existing data on the type of irrigation systems and the irrigation sources were gathered. This was done in collaboration with National Irrigation Administration (NIA). The water qualities that were gathered were compared to the existing standards of the Department of Environment and Natural Resources (DENR).

2.2 *Water Sample Collection*

Representative water samples were collected in seven major rivers of Tarlac based from the data of the National Irrigation Administration (NIA) and the Department of Environment and Natural Resources and the collection was done from 9:00 AM in the morning until 4:00 PM in the afternoon. A total of six liters of water samples were collected in each sampling sites based from the recommendation of the Department of Science and Technology. The water sampling collection was done on the onset of 2018 dry and wet season productions of rice.

2.3 *Water Quality Analysis*

Collected water samples were analyzed for its physical, chemical and microbiological qualities (Total suspended solids, chemical oxygen demand, total coliform bacteria, *E. coli*, lead and mercury content). These parameters were analyzed using the standard methods in analysis of water samples. Portable instruments were used for the analysis of the following parameters such as dissolved oxygen (portable oxygen meter), pH (HM pH-200) total dissolved solids and electrical conductivity (HM COM-100). For the nitrate quantification a Horiba portable nitrate meter was used.

2.4 *Analysis of Data*

Laboratory results from the collected water samples were analyzed and compared with the Water Quality Guidelines and General Effluent Standards of 2016 based on the Department of Environment and Natural resources (DENR) Administrative Order No. 08 Series of 2016.

edmarfranquera123@gmail.com

3 Results and Discussions

See Table 1.

3.1 *Total Soluble Solids and Chemical Oxygen Demand*

Table 2 presents the data of the different major rivers of Tarlac in terms of the total soluble solids and chemical oxygen demand. Results showed that the different river water has a varied total suspended solids and chemical oxygen demand. Higher TSS was found to be during wet season as compared during the dry season. This was also evident in terms of the chemical oxygen demand except for the two rivers, the Rio Chico and the Camiling river which exhibited a lower COD during the wet season with less than. For the TSS, based from the standard water qualifications, Tarlac and Concepcion rivers exceeded the numerical value which a body of water could be classified ranging only from 25 to 110 but for the two rivers it has both 169 mg/l total suspended solids during the wet season. Higher COD was found both in dry and wet seasons in Benig river with 27 and 22 mg/l respectively. Result of the COD laboratory test from the Benig river was also in consonance with the result of research conducted by Fernandez and David (2008) [5] which also shows high COD in Benig River. This implies that the higher COD in the sampling area, the higher level of water pollution. The wastewater discharge coming from the different industries within the area such as the presence of piggery farms could contribute to the higher COD of the water samples which maybe contributed to the deterioration of water quality within the sampling area (Al-Badaii et al. 2013).

3.2 *Heavy Metals (Lead and Mercury)*

The heavy metal concentrations (lead and mercury) in the different major rivers of Tarlac are presented in Table 3. All of the major rivers have a less than 0.05 mg/l lead and 0.0002 mg/l mercury based from the result of the laboratory analysis. Compared to the standards for the water quality the result both of the lead and mercury content of all the major rivers showed lesser than that of the standards. This implies that the rivers were not contaminated with heavy metals. This could be due to the non-presence of mining sites within the areas where the different rivers were located. Heavy metals were considered to be toxic and dangerous. The presence of higher concentrations of heavy metals in rivers as source of irrigation for the crops could lead also to the decline in production and these heavy metals could bio accumulate affecting also the humans whom will consume the crops irrigated with higher concentrations of heavy

[5] www.bgr.bund.de.Veranstaltungen. Last accessed 15 Dec 2017.

edmarfranquera123@gmail.com

Table 1 Water quality guidelines (DENR AO 08 Series 2016)

Parameter	Water body qualifications								
	AA	A	B	C	D	SA	SB	SC	SD
Dissolved oxygen (mg/l)	5	5	5	5	2	6	6	5	2
Fecal coliform (MPN/100 ml)	<1.1	<1.1	100	200	400	<1.1	100	200	400
Nitrate (mg/l)	7	7	7	7	15	10	10	10	15
pH	6.5–8.5	6.5–8.5	6.5–8.5	6.5–9.0	6.5–9.0	7.0–8.5	7.0–8.5	6.5–8.5	6.5–9.0
TSS	25	50	65	80	110	25	50	80	110
Lead (mg/l)	0.01	0.01	0.01	0.05	0.1	0.01	0.01	0.05	0.01
Mercury (mg/l)	0.001	0.001	0.001	0.002	0.004	0.001	0.001	0.002	0.004

edmarfranquera123@gmail.com

Table 2 Total soluble solids and chemical oxygen demand data of different major rivers of Tarlac province, Philippines during wet and dry season of 2018

River	Total suspended solids (mg/l)		Chemical oxygen demand (mg/l)	
	Dry season	Wet season	Dry season	Wet season
Benig	32	40	27	22
Tarlac	40	169	10	14
Bamban	58	32	11	15
Concepcion	52	169	21	19
Lapaz	223	91	11	28
Rio Chico	103	66	10	<10
Camiling	17	45	6.9	<10

Table 3 Heavy metals concentration of different major rivers of Tarlac province, Philippines during wet and dry season of 2018

River	Lead (mg/l)		Mercury (mg/l)	
	Dry season	Wet season	Dry season	Wet season
Benig	<0.05	<0.05	<0.0002	<0.0002
Tarlac	<0.05	<0.05	<0.0002	<0.0002
Bamban	<0.05	<0.05	<0.0002	<0.0002
Concepcion	<0.05	<0.05	<0.0002	<0.0002
Lapaz	<0.05	<0.05	<0.0002	<0.0002
Rio Chico	<0.05	<0.05	<0.0002	<0.0002
Camiling	<0.05	<0.05	<0.0002	<0.0002

metals. When crops were irrigated with water contaminated with heavy metals, the soils will also be polluted (Verma and Dwivedi 2013).

3.3 Dissolved Oxygen and pH

Table 4 presents the data on the dissolved oxygen and pH of the different major rivers of Tarlac province Philippines. Based from the result the highest dissolved oxygen was found to be within the Tarlac River both during the dry and wet season with 16.0 and 14.8 mg/l respectively.

The lowest dissolved oxygen was found in Concepcion River during the dry season (5.0 mg/l) and in Rio Chico River (4.8 mg/l) during the wet season. Comparing with the National standards from the DENR the major rivers of Tarlac surpasses the minimum standards of classification of water bodies with dissolved oxygen ranging from 2 to 6 mg/l. Low DO is also caused by fertilizer and manure runoff from streets, lawns and farms. The growth of too much algae which could be due to the overuse of fertilizers and the presence of fecal matters causes the speeding up of using the

edmarfranquera123@gmail.com

Table 4 Dissolve oxygen and pH of different major rivers of Tarlac province, Philippines during wet and dry season of 2018

River	Dissolved oxygen (mg/l)		pH	
	Dry season	Wet season	Dry season	Wet season
Benig	5.3	5.4	8.0	8.26
Tarlac	16.0	14.8	8.1	8.29
Bamban	9.2	6.0	8.0	7.96
Concepcion	5.0	5.0	7.0	6.78
Lapaz	8.0	5.0	7.2	7.98
Rio Chico	7.9	4.8	7.3	7.96
Camiling	15.0	14.0	8.0	8.26

oxygen quickly resulting to a lower DO.[6] The dissolved oxygen which drops below 5.0 mg/l causes stress to many aquatic lives. However based from the results, all of the rivers surpass or equal to 5.0 mg/l except for the Rio Chico River during the wet season with 4.8 mg/l.[7] In terms of pH, the major rivers of Tarlac are within the minimum and maximum standard of pH range within the DENR standards. The pH ranges from 6.78 to 8.29 during the wet season and 7.0–8.1 during the dry season.

3.4 Total Dissolved Solids and Electrical Conductivity

Higher total dissolved solids were observed in the different rivers during the dry season which ranges from 300 to 560 mg/l as compared during the wet season which ranges from 169 to 540 mg/l respectively. Too high or too low concentrations of TDS may limit the growth and may lead to the death of many aquatic organisms.[8] The reduction of water clarity, which contributes to a decrease in photosynthesis and lead to an increase in water temperature, could be due to the high concentrations of TDS. The EC during the dry season ranges from 389 to 423 while during the wet season it ranges from 280 to 420 respectively (Table 5).

3.5 Nitrate

The nitrate concentrations of the different rivers in Tarlac shows to be within the range indicated in Table 1. During the dry season, the nitrate concentrations from

[6] http://www.ririvers.org/wsp/CLASS_3/DissolvedOxygen.htm. Last accessed 30 Nov 2017.

[7] http://www.mymobilebay.com/stationdata/whatisDO.htm. Last accessed 30 Nov 2017.

[8] http://www.ei.lehigh.edu/envirosci/watershed/wq/wqbackground/tdsbg.html. Last accessed 15 Dec 2017.

edmarfranquera123@gmail.com

Table 5 Total dissolved solids and electrical conductivity of different major rivers of Tarlac province, Philippines during wet and dry season of 2018

River	Total dissolved solids (mg/l)		Electrical conductivity (μS)	
	Dry season	Wet season	Dry season	Wet season
Benig	323	218	400	323
Tarlac	308	169	420	416
Bamban	300	254	418	375
Concepcion	560	540	423	420
Lapaz	300	220	400	291
Rio Chico	305	250	412	281
Camiling	320	200	389	280

Table 6 Nitrate content of different major rivers of Tarlac province, Philippines during wet and dry season of 2018

River	Nitrate (mg/l)	
	Dry season	Wet season
Benig	14	59
Tarlac	10	48
Bamban	10	17
Concepcion	10	48
Lapaz	14	38
Rio Chico	10	45
Camiling	10	38

the different major rivers had a range of 10–14 mg/l. While during the dry season, it ranges from 17 to 59 mg/l with Benig River as the highest. The higher nutrient concentrations within the area could be due to the wastewater from the swine farm lagoons which may be discharged from the nearby farms within the area. Less than 5 mg/l N has little effect, even on nitrogen sensitive crops, but may stimulate nuisance growth of algae and aquatic plants in streams, lakes, canals and drainage ditches (Table 6).[9]

3.6 *Fecal Coliform and* E. coli

In terms of the microbiological parameters such as fecal coliforms and *E. coli*, the different river waters of Tarlac was higher than the standards particularly in Benig River with 11,000 MPN/100 ml and within the Concepcion river which exceeds the National standards for safe water with fecal coliform count of 140,000. Higher concentrations of *E. coli* were also noted in Benig and Concepcion River both with

[9]http://www.fao.org/docrep/003/T0234E/T0234E06.htm. Last accessed 15 Dec 2017.

edmarfranquera123@gmail.com

Table 7 Fecal coliform and *E. coli* concentration of different major rivers of Tarlac province, Philippines during wet and dry season of 2018

River	Fecal coliform (MPN/100 ml)	*E. coli* (MPN/100 ml)
	Wet season	Wet season
Benig	11,000	1700
Tarlac	390	21
Bamban	270	17
Concepcion	140,000	1700
Lapaz	2600	170
Rio Chico	2800	330
Camiling	330	<1.8

1700 MPN/100 ml. The high concentrations within the said rivers could be due to the wastewater discharged from the nearby areas contributing to the higher Fecal coliform and *E. coli* in the said areas of concern. The higher concentrations as observed in the two rivers could have a potential to reduce the water quality thus reducing also the recreational value (Table 7).[10]

4 Conclusions

The water samples collected from major rivers of Tarlac revealed that there were variations in the results in terms of the different parameters used to quantify the concentrations of the physical, chemical and microbiological quality of the river waters for irrigation purposes. Based from the result, the different river waters were also in accordance with the National Standards set by the Department of Environment and Natural Resources (DENR).

Acknowledgements The authors would like to acknowledge the Department of Agriculture Regional Field Office (DA-RFO 3) for funding this research and sincere acknowledgement was also to the management of the Tarlac Agricultural University.

References

Al-Badaii F, Shuhaimi-Othman M, Gasim MB (2013) Water quality assessment of the Semenyih River, Selangor, Malaysia. J Chem Article ID 871056, 10 p. https://doi.org/10.1155/2013/871056

Fernandez XD, David ME (2008) Water quality assessment of the benig river: implication to environmental management accessed through https://www.bgr.bund.de/EN/Themen/Wasser/Veranstaltungen/symp_sanitat-gwprotect/poster_fernandez_pdf.pdf?__blob=publicationFile&v=2 December 2017

[10]https://pubs.usgs.gov/wri/wri004139/pdf/wrir00-4139.pdf. Last accessed 15 Dec 2017.

edmarfranquera123@gmail.com

Verma R, Dwivedi P (2013) Heavy metal water pollution—a case study. Recent Res Sci Technol 5(5):98–99. ISSN: 2076-5061. Available Online http://recent-science.com/

edmarfranquera123@gmail.com

International Journal on Advanced Science Engineering Information Technology

Vol.9 (2019) No. 3
ISSN: 2088-5334

Enhanced Manhattan-based Clustering using Fuzzy C-Means Algorithm for High Dimensional Datasets

Joven A. Tolentino[#, *], Bobby D.Gerardo[#]

[#]*Technological Institute of the Philippines, Quezon City, Philippines*
E-mail: jatolentino@tau.edu.ph; bobby.gerardo@gmail.com

[*]*Tarlac Agricultural University, Tarlac, Philippines*

***Abstract*—The problem of mining a high dimensional data includes a high computational cost, a high dimensional dataset composed of thousands of attribute and or instances. The efficiency of an algorithm, specifically, its speed is oftentimes sacrificed when this kind of dataset is supplied to the algorithm. Fuzzy C-Means algorithm is one which suffers from this problem. This clustering algorithm requires high computational resources as it processes whether low or high dimensional data. Netflix data rating, small round blue cell tumors (SRBCTs) and Colon Cancer (52,308, and 2,000 of attributes and 1500, 83 and 62 of instances respectively) dataset were identified as a high dimensional dataset. As such, the Manhattan distance measure employing the trigonometric function was used to enhance the fuzzy c-means algorithm. Results show an increase on the efficiency of processing large amount of data using the Netflix ,Colon cancer and SRCBT an (39,296, 38,952 and 85,774 milliseconds to complete the different clusters, respectively) average of 54,674 milliseconds while Manhattan distance measure took an average of (36,858, 36,501 and 82,86 milliseconds, respectively) 52,703 milliseconds for the entire dataset to cluster. On the other hand, the enhanced Manhattan distance measure took (33,216, 32,368 and 81,125 milliseconds, respectively) 48,903 seconds on clustering the datasets. Given the said result, the enhanced Manhattan distance measure is 11% more efficient compared to Euclidean distance measure and 7% more efficient than the Manhattan distance measure respectively.**

***Keywords*— fuzzy C-Means; high dimensional dataset; Manhattan distance; clustering.**

I. INTRODUCTION

The high dimensional dataset is common nowadays due to the colossal amount of information being gathered electronically by varying information systems. Movies, medical health record, and agricultural dataset can be observed to be as high dimensional dataset. Duplication of records, multiple attributes and thousands number of records were categorized as high dimensional datasets, and most of the data mining algorithms suffer low accuracy and high computational cost in processing when a high dimensional dataset was supplied [1]. This high dimensional dataset can also be observed to know what this dataset shows and implies.

A common technique to observe this dataset is using clustering. Clustering splits a large amount of data and performs grouping considering the similarities of the individual data supplied [2]. However, several clustering algorithms suffer from high computational cost and one of which is the Fuzzy C-Means algorithm.

Fuzzy C-Means also suffers from its accuracy and speed when a dataset contains high dimension or not [3], [4]. The study aims to enhance the Fuzzy C-Means algorithm by changing the distance measure to solve the weakness of the said algorithm. Manhattan distance measure was used since it is also ideal when applied to high dimensional dataset [5]. The trigonometric approach was utilized to the said distance measure since the accuracy of the Manhattan distance measure suffers when centroid and points are connected diagonally [6], [7].

Data mining procedures will also be used to prepare the actual dataset for mining. The computational cost will be observed by testing the algorithm with different distance measures (Euclidean, Manhattan and Enhanced Manhattan) and three different high dimensional datasets (Netflix Movie Rating, Colon Cancer and SRCBT) which will lead on what specific distance measure is faster when applied to the said algorithm.

II. MATERIALS AND METHOD

To investigate the performance of the modified algorithm, Knowledge Discovery Model were used proposed by [8] consisting the step of data selection, data pre-processing, transformation, and data mining. Figure 1 shows the actual process of KDD.

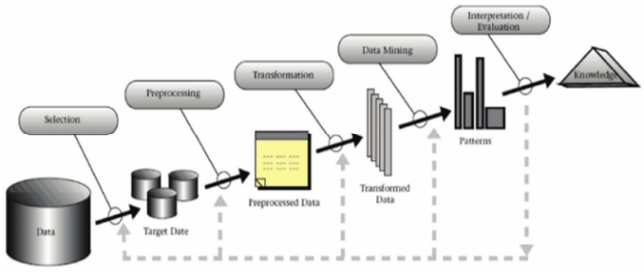

Fig.1 the Knowledge Discover Model

With the KDD model, the dataset should be ideal to be processed from the part of the selection to the step of data mining.

A. Data Selection

Selection of the actual dataset for clustering was done by searching for the appropriate dataset that has high dimensionality. High dimensional datasets are the ones who have multiple fields and thousands of records [1]. The high dimensionality of data is also when dataset features are greater than the number of instances [9]. The Netflix movie rating, small round blue cell tumors (SRBCTs) and Colon Cancer dataset are also categorized as high dimensional considering these definitions. Table 1 shows the number of features of the said datasets.

TABLE I
THE FEATURES AND INSTANCES OF THE DATASETS

Dataset	Features	Instances
Netflix Movie	5	1,500
Colon Cancer	2000	62
SRCBT	2308	83

B. Pre-Processing

The pre-processing technique was also done to prepare the dataset that will be used. This technique reduces the dimensionality of the dataset [10], [11]. The dataset was merged into a file and field were also observed to identify the process needed to be done to reduce its dimensionality. The term discretization technique describes another advantage of this step. In this part, the equal frequency binning was used. This step converts the text into a numeric value. Each instance in the dataset that has the same value are considered as one and converted to a similar numeric value [12]. In Table II, the values of the feature, genre, were discretized to fit the algorithm.

TABLE II
A PORTION OF THE MOVIE DATASET WITH ITS GENRE

No	Movie Title	Genre	Discretised Value
1	Cat Run 2 (2014)	Action	1
2	He Who Dares (2014)	Action	1
3	How to Train Your Dragon 2 (2014)	Action|Adventure|Animation	2
4	Hercules (2014)	Action|Adventure	3
5	Falcon Rising (2014)	Action|Adventure	3
6	Land Ho! (2014)	Adventure|Comedy|Documentary|Mystery	4
8	Seventh Son (2014)	Adventure|Children|Sci-Fi	6

In this process, the field, genre, was discretized to be applicable with clustering. The same process was done for the two remaining datasets (Colon Cancer and SRCBT). The feature class was converted into a numeric value.

C. Transformation

Making the dataset suitable for knowledge discovery requires the dataset to be transformed. The dataset for Netflix movie rating is composed of several tables (Movie, Rating, and Tags) that are connected via Primary Key (PK) and a Foreign Key (FK). A foreign key is several techniques can do a specific property of dataset, which is described by the implementation of the primary key to another data table [13] and merging this dataset. One technique for combining this data table for preparation for data mining is union. The union is the process of identifying the intersection of two or more data table with their PK and FK[14]. Hence, the researcher created a tool for merging the data table into a single dataset concerning the primary key and foreign key.

For the two remaining datasets, features were already normalized aside from the pre-processing technique. Observation of the actual content of the dataset was also needed to be observed thoroughly to see how these datasets were constructed such that the enhanced algorithm can process it. Based on the pre-processing and transformation techniques, the following portions of the datasets of Netflix Movie Rating, Colon Cancer and SRCBT had been derived.

TABLE III
NETFLIX MOVIE DATASET

Rating	UserID	Time Stamp	Genre	MovieID
2.5	53930	1393064439	30	22306
2.5	87813	1387131563	30	22306
3	137200	1398867354	30	22306
4.5	13494	1421295240	114	23623
4	15720	1426647292	114	23623

TABLE IV
COLON CANCER DATASET

FTR1	FTR1	FTR1	FTR to 2000	Class
88.23	39.67	67.83	28.7	2
82.24	85.03	152.2	16.77	1
76.97	224.62	31.23	15.16	2
74.53	67.71	48.34	16.09	1
54.56	223.36	73.1	31.81	2
33.2	91.85	5.88	21.88	1
98.54	54.62	30.54	24.45	2

TABLE V
SRCBT DATASET

FTR1	FTR1	FTR1	FTR to 2308	Class
0.143	0.888	0.068	0.108	2
0.085	0.324	0.635	0.271	1
0.193	0.39	0.378	0.107	3
0.159	0.248	1.164	0.224	4

D. Data Mining

Clustering algorithm will be enforced in this study by using the Fuzzy C-means algorithm. This tool can be used to address its problem on clustering high dimensional datasets. Figure 2 shows the actual process of how Fuzzy C-Means Clustering works.

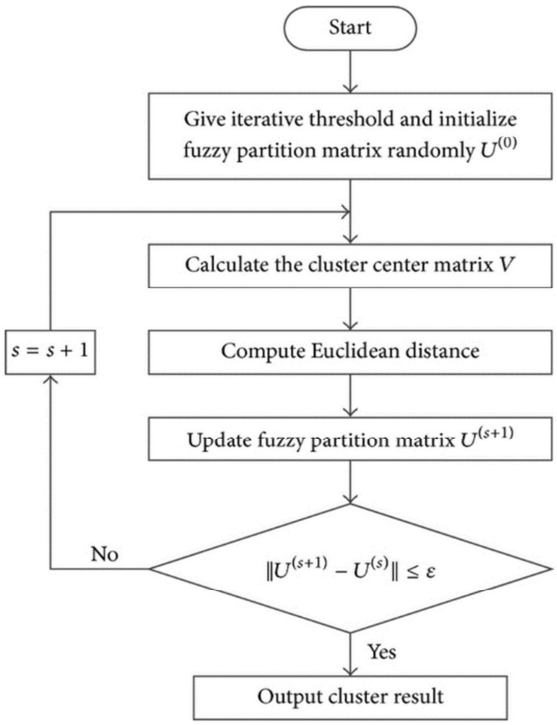

Fig. 2 The actual process of clustering using Fuzzy C-means algorithm

The first step is that Fuzzy C-Means selects the number of cluster and membership functions ranging from zero to one. The calculation of the actual centroid with the corresponding parameter follows. The computation of the actual centroid plays a vital role in creating the clusters[15]. This will identify how many iterations will be done. The third step is to date the actual cluster with the specific distance measure and lastly, validate the result. The iterations take place until convergence is achieved [16]. With this given process of Fuzzy C-Means algorithm, changing the distance measure can improve the performance of the said algorithm.

E. Manhattan Distance Measure

Providing a result with lesser computational cost can be achieved using different strategies. Observing the distance measure used by the algorithm and its performance can be a basis in identifying what distance measure is applicable for the high dimensional dataset. The Manhattan distance measure is commonly used when the point that is generated was vertically or horizontally connected. Selecting an appropriate distance measure plays a vital role in providing a good set of clusters [17]. The study also shows that Manhattan distance measure is more accurate in the calculating distance when the dataset is high dimensional compared to other distance measures [18]. Table VI shows the side by side comparison of several distance measure.

TABLE VI
COMPARISON OF SEVERAL DISTANCE MEASURE.

Distance Measure	Benefits	Drawbacks
Euclidean	Easy to Implement and Test	Results are greatly influenced by variables that have the largest value. Does not work well for Image data, Document Classification
Manhattan	Easily generalized to a higher dimension	Does not work well for image data and document classification
Cosine	Handles both continuous and categorical variables	Does not work well for nominal data
Jaccard Index	Handles both continuous and categorical variables	Does not work well for nominal data

The use of the Manhattan Distance Measure will allow the algorithm to speed up its processing time, although Manhattan distance measure has a problem needed to be addressed.

F. Euclidean Distance Measure

On the other hand by default Euclidean distance measure were used in Fuzzy C-Means, it produces a more accurate result but higher computational cost [19], this is the main reason why the algorithm needed to be improved with the proposed modification conceptualized.

G. Enhancement of the Distance measure

A weakness of the Manhattan distance measure is in terms of clustering points that are connected diagonally. Fig. 3 shows the actual points connecting to the centroid diagonally.

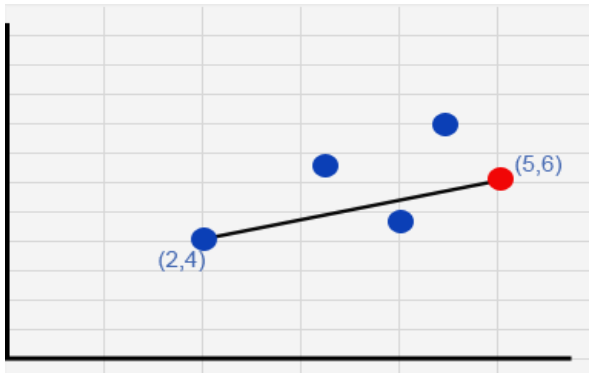

Fig. 3 Centroid and data point are connected diagonally

Employing trigonometric function specifically COSINE Equation 2 was tested in order to check the speed of the said

distance measure when applied to the Fuzzy C-Means algorithm.

$$cosine(\emptyset) = \frac{\text{adjacent}}{hupothenuse} \quad (2)$$

Where adjacent (next to) is to the angle θ and Hypotenuse is the long line, equation 3 shows the actual solution to address the problem of Manhattan Distance.

$$d = \sum \frac{|(x3-x2)+ (y_2-y_3)|}{cosine(\emptyset)} \quad (3)$$

Where y_3 is the point of intersection from the created imaginary line and y_2 is the y-coordinate of the centroids. The difference of y_3 and y_2 will be divided to $cosine(\emptyset)$. $Theta$ $(\emptyset)$ is used since the actual angle is not yet solved. To calculate the actual distance the following, steps were considered.

Step 1. *Create an imaginary line to form a right triangle*

Step 2. *Identify the point of intersection*

Step 3. *Compute the Distance of the Imaginary line using Manhattan. Given that (x2=5,x3=5) and (y2=6,y3=4)*

(5-5)+(6-4)= 2

Step 4. *Compute for the distance*

2/Cosine(53.60)=3.61

The given steps in calculating the actual distance of the centroid to the dataset points may lead to higher accuracy for the Fuzzy C-Means Algorithm when supplied.

H. Fuzzy C-Means

To further test the algorithm, the steps for the distance measure were invoked by the enhanced Manhattan distance measure. By default, Fuzzy C-Means uses Euclidean distance. Fig. 4 shows the actual process of clustering the dataset using the enhanced Manhattan distance.

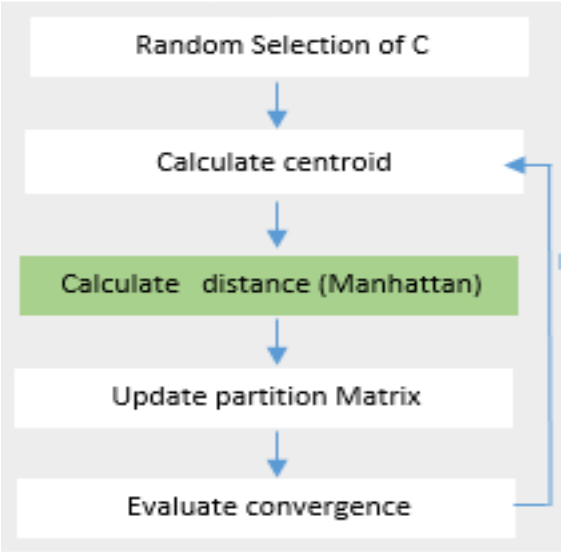

Fig. 4 The actual process of clustering using Fuzzy c-means algorithm.

The following pseudo-code was used to implement the Modified Manhattan distance measure over Fuzzy C-Means Algorithm.

Start
Required Array of Points and Centroid
Declare Distance

For counter=0; to LengthofPoints step 2

If Centroid is equal to Points
Get the absolute difference of points and the centroid
Else
Get the absolute difference of centroid and Imaginary line
Divide the absolute difference to $cosine(\emptyset)$

EndIf

Update distance by adding the difference
Iterate to each column and pair it with (x,y) format and do the calculation for the distance.
End

I. Evaluation

To validate the performance of the said modified algorithm, the duration to complete the process of clustering using Fuzzy C-Means with different distance measures were compared, and the starting points of the program were tracked. The differences were calculated to identify how many milliseconds were needed to complete the actual clustering process. The following pseudo code was used to evaluate the performance of Fuzzy C-Means on applying the three distance measures and three high dimensional datasets.

Start
Get Start time in milliseconds
Declare Threshold=1, iteration=0
While Threshold is not equal to 0
Update value of iteration +1
Assign new center
End While
Elapsed time = end time - start time
End

The process of Fuzzy C-Means clustering stops when the convergence is reached. This means that when the threshold becomes zero, the actual clustering process is finished on clustering [16] and as prescribed by the algorithm threshold use was zero. Tracking the execution time of the program can now be observed along with the behavior of the algorithm when different distance measures and different datasets with high dimensions was applied.

III. Results and Discussion

With the procedure of pre-processing and transformation, the dataset Netflix Movie composed of 4 attributes with 1,500 instances, Colon Cancer having 2000 features and 65 instances and SRCBT 2308 features and 83 instances are

now ready for clustering and comparison of the actual speed of the modified algorithm to the standard Fuzzy C-Means Algorithm. The algorithm was tested by computing the actual time elapsed when the clustering processes were simulated. Table VI showed the actual result of the algorithm when Euclidean and Enhanced Manhattan distance measures were used.

TABLE VII
RESULT OF THE ALGORITHM IN (MS), WHEN EUCLIDEAN AND ENHANCED MANHATTAN IS USED

Dataset	Clusters	Euclidean	Enhanced Manhattan
Netflix Movie	10	39296	33216
Cancer	4	38952	32368
SRCBT	3	85774	81125

Observing the actual result, the Enhanced Manhattan distance measure outperformed the Euclidean distance Measure. To further investigate, the Manhattan distance measure was also used to compare the actual results as shown in Table VIII.

TABLE VIII
RESULT OF THE ALGORITHM IN (MS), WHEN MANHATTAN AND ENHANCED MANHATTAN IS USED

Dataset	Clusters	Manhattan	Enhanced Manhattan
Netflix Movie	10	36858	33216
Cancer	4	36501	32368
SRCBT	3	82860	81125

With the dataset supplied to the Manhattan distance and enhanced Manhattan distance, the result shows that the actual modification decreases the processing time for clustering the three datasets. Comparison of the actual result for clustering using Fuzzy C-Means with the three distance measure is indicated in Figure 5 and 6. The behavior of the algorithm varies on the dataset supplied, especially when it comes to high dimensions.

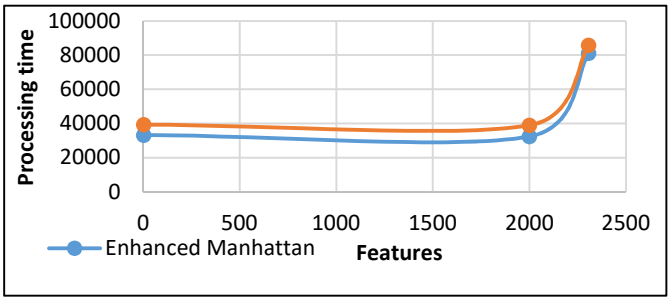

Fig. 5 The comparison of the processing time of Euclidean and Enhanced Manhattan against the Dataset Features.

The trend of the three distance measure plotted along with the number of features and to its processing time showed an improvement when the Enhanced Manhattan Distance measure was supplied. This indicates that the modification can now be applied to the algorithm to increase its speed on clustering high dimensional datasets.

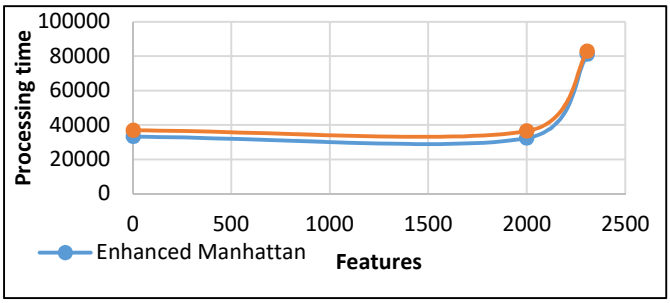

Fig. 6 The comparison of the processing time of Manhattan and Enhanced Manhattan against the Dataset Features.

The trend of the three distance measure plotted along with the number of features and to its processing time showed an improvement when the Enhanced Manhattan Distance measure was supplied. This indicates that the modification can now be applied to the algorithm to increase its speed on clustering high dimensional datasets.

IV. CONCLUSION

Fuzzy C-means algorithm is an algorithm that suffers from high computational cost when a high dimensional dataset is applied. One way to address the said problem is by invoking the distance measure used. In this study, an Enhanced Manhattan-based clustering was used employing trigonometric function to address the issue of Manhattan distance measure.

Results show that an increase in the efficiency in terms of speed of the said algorithm can be observed when using the enhanced Manhattan distance measure. Euclidean distance measure shows that clustering the three datasets such as Netflix Movie Rating, Colon Cancer, and SRBT has a (39,296, 38,952 and 85,774 milliseconds to complete the different clusters, respectively) average of 54,674 milliseconds while Manhattan distance measure took an average of (36,858, 36,501 and 82,86 milliseconds, respectively) 52,703 milliseconds for the entire dataset to cluster. On the other hand, the enhanced Manhattan distance measure took (33,216, 32,368 and 81,125 milliseconds, respectively) 48,903 seconds on clustering the datasets.

Given the said result, the enhanced Manhattan distance measure is 11% more efficient compared to Euclidean distance measure and 7% more efficient than the Manhattan distance measure respectively. While the efficiency increases for the said algorithm, it needs further observation on the behavior of the algorithm in clustering a standard type of dataset. Accuracy also needs to be studied in applying this modified algorithm. Other factors can also be considered to increase the efficiency of the said algorithm.

ACKNOWLEDGMENT

This study would not be possible without the support of the Commission on Higher Education Kto12 Transition Program Unit - Quezon City, Philippines, and the Tarlac Agricultural University Tarlac, Philippines. Gratitude is also extended to the Technological Institute of the Philippines – Quezon, City.

REFERENCES

[1] N. Raksha and R. Alankar, "Detection of fuzzy duplicates in high dimensional datasets," *2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI, 2016*, pp. 1423–1428, 2016.

[2] Y. G. Jung, M. S. Kang, and J. Heo, "Clustering performance comparison using K- means and expectation maximization algorithms," *Biotechnol. Biotechnol. Equip. ISSN1310-2818*, vol. 2818, no. October 2015.

[3] Z. Cebeci and F. Yildiz, "Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures," *J. Agric. Informatics*, vol. 6, no. 3, pp. 13–23, 2015.

[4] R. Winkler, F. Klawonn, and R. Kruse, "Problems of Fuzzy c-Means Clustering and Similar Algorithms with High Dimensional Data Sets," *Challenges Interface Data Anal. Comput. Sci. Optim.*, pp. 1–8, 2012.

[5] S. Pandit and S. Gupta, "A Comparative Study On Distance Measuring," *Int. J. Res. Comput. Sci.*, vol. 2, no. 1, pp. 29–31, 2011.

[6] T. K. Mohana, V. Lalitha, L. Kusuma, N. Rahul, and M. Mohan, "Various Distance Metric Methods for Query Based Image Retrieval," vol. 7, no. 3, pp. 5818–5821, 2017.

[7] M. Khan and T. Shah, "A copyright protection using watermarking scheme based on nonlinear permutation and its quality metrics," *Neural Comput. Appl.*, vol. 26, no. 4, pp. 845–855, 2014.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in," vol. 17, no. 3, pp. 37–54, 1996.

[9] J. Zhang and M. Pan, "A high-dimension two-sample test for the mean using a cluster," *Comput. Stat. Data Anal.*, vol. 97, pp. 87–97, 2016.

[10] L. Zhou, "Preprocessing Method before Data Compression of Cloud Platform," pp. 1223–1227, 2017.

[11] M. A. Chaudhari, P. M. Phadatare, P. S. Kudale, R. B. Mohite, and R. P. Petare, "Preprocessing of High Dimensional Dataset for Developing Expert IR System," pp. 417–421, 2015.

[12] Z. Marzuki and F. Ahmad, "Data Mining Discretization Methods and Performances Data Mining Discretization Methods and Performances," no. December, pp. 3–6, 2014.

[13] N. A. Mian and N. A. Zafar, "Key Analysis of Normalization Process using Formal Techniques in DBRE," 2010.

[14] C. Ordonez, "Data Set Preprocessing and Transformation in a Database System," vol. 15, no. 4, pp. 1–19, 2011.

[15] Z. Wang, N. Zhao, W. Wang, R. Tang, and S. Li, "A Fault Diagnosis Approach for Gas Turbine Exhaust Gas Temperature Based on Fuzzy C-Means Clustering and Support Vector Machine," *Math. Probl. Eng.*, vol. 2015, pp. 1–11, 2015.

[16] N. Grover, "A study of various Fuzzy Clustering Algorithms," *Int. J. Eng. Res.*, vol. 5013, no. 3, pp. 177–181, 2014.

[17] L. H. Son, "Generalized picture distance measure and applications to picture fuzzy clustering," *Appl. Soft Comput. J.*, pp. 1–12, 2016.

[18] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," pp. 420–434, 2001.

[19] A. Fahad *et al.*, "A Survey of Clustering Algorithms for Big Data : Taxonomy & Empirical Analysis," 2014.

Modified Graph-theoretic Clustering Algorithm for Mining International Linkages of Philippine Higher Education Institutions

Sheila R. Lingaya[1], Bobby D. Gerardo[2], Ruji P. Medina[3]
Technological Institute of the Philippines - Quezon City, Philippines[1, 3]
Tarlac Agricultural University[1]
West Visayas State University[2]

***Abstract*—Graph-theoretic clustering either uses limited neighborhood or construction of a minimum spanning tree to aid the clustering process. The latter is challenged by the need to identify and consequently eliminate inconsistent edges to achieve final clusters, detect outliers and partition substantially. This work focused on mining the data of the International Linkages of Philippine Higher Education Institutions by employing a modified graph-theoretic clustering algorithm with which the Prim's Minimum Spanning Tree algorithm was used to construct a minimum spanning tree for the internationalization dataset infusing the properties of a small world network. Such properties are invoked by the computation of local clustering coefficient for the data elements in the limited neighborhood of data points established using the von Neumann Neighborhood. The overall result of the cluster validation using the Silhouette Index with a score of .69 indicates that there is an acceptable structure found in the clustering result – hence, a potential of the modified MST-based clustering algorithm. The Silhouette per cluster with .75 being the least score means that each cluster derived for r=5 by the von Neumann Neighborhood has a strong clustering structure.**

***Keywords*—MST-based clustering; Small World Network; von Neumann Neighborhood; internationalization; Prim's MST**

I. Introduction

Internationalization and partnership development undertakings pave way to establish identity in the international arena. As such, data in the field of internationalization as mirrored by students and international partnerships established by education institutions is growing to be a good interest of researches [1]–[4]. This is since the rate of internationalization increases with the unhindered channels of communications and affordable travel expenses. Universities seek to seize the opportunities from global partnerships and foster relationships with other organizations or institutions. Internationalization is also described to transform into mainstream strategy in higher educations and is increasingly seen as adding value to the life of universities through improving their quality [5]. The definition of internationalization being the process of integrating international, intercultural, or global dimensions into the purpose, functions or delivery of post-secondary education [6] is by common knowledge, the most frequently cited and widely accepted.

Meanwhile, methods and techniques in data mining allow analysis of very large datasets (i.e. big data) to extract and discover previously unknown structures and relations out of huge amount of details [7] for the purpose of knowledge extraction. As such, clustering in the data mining arena aims to establish high intra-cluster and low inter-cluster similarity in data. The high intra-cluster similarity should be based on the derived measurement from the data while the low inter-cluster similarity should maintain that elements in the different clusters should have maximum distance. These are intended to achieve beneficial knowledge from the data [8] for decision making and strategizing. Among different types of clustering, the most conventional distinction is whether the set of clusters is hierarchical or partitional [9] where hierarchical is a set of nested clusters while partitional clustering divides the set of data objects into non-overlapping clusters such that each object is in exactly a single cluster [10]. However in the real world, clusters come in arbitrary shapes, varied densities and unbalanced sizes that is why there is no universal clustering method which can deal with all problems [11].

Since most clustering algorithms' performance is affected by the shape and size of the detectable clusters [12], the requirement of an *a priori* knowledge about the actual number of clusters and the setting of a threshold to obtain adequate clustering results; a number of modifications to the clustering algorithms have emerged and are being explored to cope with said problems. Among which are graph-theoretic or graph-based clustering algorithms where data is represented in an undirected graph denoted as G={V,E} where the set of all data points is V and the set of connections between two distinct data objects (i.e. edges or links are contained in E. This is associated with a distance measure resulting to a connected subgraph or clusters. The use of Minimum Spanning Tree (MST) is one of said methods which either uses the Prim's [13]–[15] or Kruskal's [16], [17]. An MST is constructed for the whole data with a threshold value being set along with a number of steps to terminate the process to form clusters resulting from removing an inconsistent edge whose value is greater than the threshold value. However, this strategy is constrained by the identification and elimination of the inconsistent edge [17], detection of outliers [18], as well as insufficiently evidenced partitioning–hence, having the same weaknesses as other clustering methods that are based on distance measures [19].

This work aims to perform data mining in the data of the international linkages of Philippine Higher Education Institutions (PHEIs) using a proposed modified Prim's MST-based clustering algorithm producing a minimum spanning tree for the dataset infusing the computation of local clustering coefficient for the data points in the limited neighborhood generated by von Neumann Neighborhood.

This paper is organized as follows. Section II presents the conceptual framework of the modified Prim's MST-based clustering algorithm invoking the properties of the small-world network of graph theory. It also highlights the preparation of the International Linkages data. Section III includes the results of the simulation and the cluster validation. Section IV highlights the conclusions and future works of the study.

II. Modified Prim's MST-based Clustering Algorithm

Clustering can be used on many problems as it is helpful to seek and see relationships. It aims to congregate into clusters unlabeled data elements with high similarity based on a measure obtained solely from the data itself [20]. The distance measure defines the radius of membership depending on the type of data on hand. A good cluster is associated with high clustering value in terms of distance so the selection of distance metric is essential in clustering [21] while another clustering algorithm approach is to represent a target data set as a weighted undirected graph [20].

A. Prim's MST-based Clustering Algorithm

Prim's MST Algorithm uses a distance function to specify the closeness of data objects to establish the weight between them by choosing an arbitrary point to the next adjacent point of minimum weight. For clustering, an edge inconsistency measure is defined to identify an inconsistent edge to be removed to partition the whole dataset into clusters. Prim's MST is modified for efficient construction of spanning tree based on the k-nearest neighbor search mechanism during which a new edge weight is defined to maximize the intra-cluster similarity and minimize the inter-cluster similarity [13]. The algorithm can be used for a complete graph while using Fibonacci Heap [19], [22].

In this work, the traditional Prim's MST Algorithm for clustering defined by [18] as shown in Fig. 1 is modified by infusing the local neighborhood search by the von Neumann Neighborhood in order to facilitate the computation of the local clustering coefficients of the data elements in said neighborhood.

Higher clustering coefficient indicates the robustness on an average shortest path between any pair of nodes [23]–[25]. As such, small world networks [26] have the properties of having a small mean of shortest path length and high clustering coefficient. The Local Clustering Coefficient (LCC) quantifies the closeness of the neighbors of a vertex in becoming a clique. A concept in graph theory, LCC is basically computed as the number of triangles connected to a vertex over the number of triples around a given vertex. It is the probability that duos of neighbors of a vertex are connected by an immediate connection – the value is $0 \leq LCC \leq 1$. Thus,

$$LCC = \frac{number\ of\ connected\ neighbors}{number\ of\ neighbors} \quad (1)$$

Meanwhile, the von Neumann Neighborhood is one of the most commonly used types of neighborhood for cellular automata of two dimensions [27]. It is also used in pattern generation [28] and operations research [29] as it has been proven to have better performance than other topologies to further improve the quality of local search [30]. It can be extended by taking the set of data objects at Manhattan distance r where $r > 1$ which yields a result of a diamond-shaped region – hence, the neighborhood of data objects. The two-dimensional square lattice is composed of the central cell and the four adjacent cells around achieved through traversing North, East, West and South (NEWS) derived at a Manhattan distance 1. The number of neighbors (i.e. cells) in 2-dimensional by von Neumann Neighborhood of the cellular evolutionary algorithm for range r is defined as:

$$2r(r+1)+1 \quad (2)$$

As such, the modified Prim's MST-based Clustering Algorithm establishes the adjacency of the data facilitated by the suitable cellular automaton, the von Neumann Neighborhood which simulates the establishment of neighborhood. This preludes the computation of local efficiency or local clustering coefficient. Thus, the modified Prim's MST construction for clustering is defined by (u, v, LCC(v), d(u,v)) such that u is the initial data point and v is the terminal data point.

While the traditional Prim's MST considers only the next minimum distance d(u,v) between data u in the MST being built T and the adjacent data point v in V; the modified algorithm initially considers the LCC of the adjacent data point LCC(v) to ensure a high clustering coefficient for the whole cluster – hence, pursuing clusters of density. As Prim grows the MST one edge at a time, it should be noted that the next candidate edge or connection of data point must respect the partition or cut of the set of points in the minimum spanning tree T and V to avoid having a cycle.

Pseudocode for PMST for Clustering

```
procedure MST Clustering (V: set of data points v )
construct a fully connected graph G of V such that
the
   edge weights are the distances between data
points
construct  Prim's minimum spanning tree T of G
maintain disjoint sets V and T
select minimum d(v,u) where v∈VT and u∈V
check for cycle
find all inconsistent edges of T
remove inconsistent edges to get a set of connected
    components
define the connected components as clusters
```

Fig. 1. Prim's Minimum Spanning Tree for Clustering.

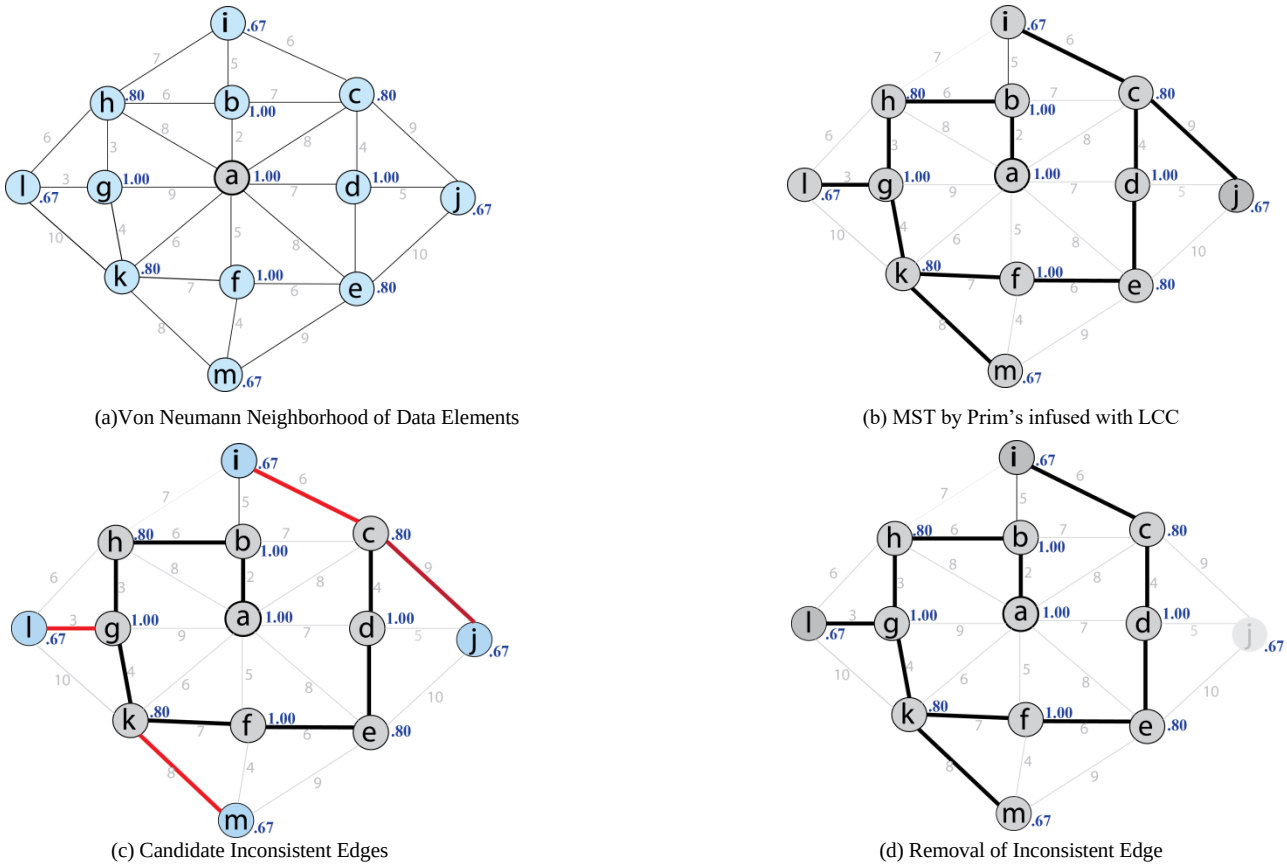

Fig. 2. Prim's Minmum Spanning Tree Construction for a Local Neighborhood Established using Von Neumann.

Being the data element having the least LCC(v) and maximum distance d is the criterion set for identifying the inconsistent edge. Data elements l, j, k and m in Fig. 2(a) are all LCC=.67 – hence, their distances to the data points in the MST were considered as indicated in Fig. 2(b). As such, d(c,j)=9 indicated in Fig. 2(c) as the connection with the greatest distance is the inconsistent edge. The algorithm will herein iterate and continue on the other data points of the data set. The resulting MST must have N-1 edges for N number of data points without cycle – hence, the cluster as seen in Fig. 2(d).

B. Data Cleaning and Preparation

The PHEI International Linkages data contains the actual and essential records for the international linkages of Philippine Higher Education Institutions. It is consist of partnerships entered by PHEIs with foreign universities and/or organizations transpiring into different internationalization activities including student exchange, faculty exchange, academic collaboration, research collaboration and other activities across different disciplines. The dataset is summarized in Table I.

An integral part of the data mining process is the data to which knowledge discovery is applied. The International Linkages data contains instances of inconsistencies, incompleteness and variations in the essence of data mining. As such, entries or values were simplified and prepared such that the proposed algorithm is able to process it. In the original data, the field for partnership form has duplicative entries and no defined options. A particular discipline is mentioned in several groups with each specific to a particular partnership. A similar case can be observed with an area of partnership (e.g. Faculty Exchange) being included and specific in a number of partnerships. Hence, in Table II are the disintegrated attributes rooted from the form of linkages attribute of the original data.

TABLE I. LIST OF PHEI LINKAGES DATASET FEATURES AND DESCRIPTION

Name	***Definition***	***Example***
country	where foreign university or organization partnered is located	Indonesia
continent	where country of foreign university or organization partnered is located	Asia
phei	the Philippine Higher Education Institution (e.g. SUC, HEI)	TAU
partner	name foreign university or organization partnered	CRRU
p_form	form of partnership	Bilateral
p_area	area of internationalization activities	Faculty Exchange
p_discipline	field of discipline covered by the partnership	Education
d_signed	date when partnership was signed	02/06
p_year	year when partnership was signed	2017
p_status	if active or inactive	Active

TABLE II. CONTENT RELATED FEATURE OF ATTRIBUTES DISINTEGRATED FROM FIELD

Name	***Definition***
p_type	Bilateral; Multilateral
p_area	Faculty Exchange; Student Exchange; Research(er) Exchange/Collaboration; Academic Collaboration; Joint Publication
p_discipline	Accounting, Arts, Education, Fashion and Textiles, Social Studies

The conversion of the textual values was necessary since most instances are texts and multiple values are specific to one entry. The data cleaning and preparation executed is where each distinct group is coded. For instance, in the area of partnership terms, the PHEI can either use its own nomenclature but certainly, it may also use the terms of reference by the prospect foreign partner university or organization. Hence, all attributes were coded and assigned a numerical value to discretize the data so that the clustering algorithm will be able to process it.

III. RESULTS AND DISCUSSION

The cluster analysis of the data on international linkages of PHEIs aimed to gain valuable insights of the data to see what groups the data elements belong to while having the modified clustering algorithm to define instances with similar properties as a group. Data may come into mix type in the real world such that one attribute may be expressed in ration and others in terms of categorical that adjustment may be hard in terms of the algorithm because some specific algorithms can only be applicable to certain types of data. There may be a need for some data transformation or preprocessing to do so that the algorithm will work. Data cleaning and preprocessing was an integral part of the data mining process to make adjustments and the data be made suitable with the proposed algorithm as it cleaned and prepared the data for the algorithm to be able to process it.

A. Simulation

The algorithm was implemented through the following Pseudocode in Fig. 3 and simulated on the discretized Internationalization data set.

```
Pseudocode for Modified Prim's MST-based Clustering with
Local Efficiency

procedure MST Clustering (V: set of data points v )
Remove all redundant data
set arbitrary data point
get arbitrary data point's Neighborhood
generate connection for each data point in neighborhood;
set LCC for each data point in neighborhood
construct  Prim's minimum spanning tree T of G
maintain disjoint sets V and T
set data point with least LCC and maximum distance as
    inconsistent edge
remove inconsistent edges to get a set of connected
    components
define the connected components as clusters
```

Fig. 3. Modified Prim's MST-based Clustering with Local Efficiency.

The International Linkages data set is composed of 12 attributes with 748 instances. With a random value r=5, seven clusters were generated. The attributes with only at most 2 possible values were not used for the experiment.

Two attributes (e.g. continent, phei) were used to define an instance-hence to illustrate, data point (x, y) defines one data element by its value on attributes continent and phei as x and y, respectively. The neighborhood of said data points determined by NEWS was derived with the nearest higher value in x for north, nearest lower value in x for south, nearest higher value in y for east, and nearest lower value in y for south until the prescribed number of neighbors of the arbitrarily chosen value through von Neumann's Neighborhood is derived.

An observation on the result of the presented data mining procedure is that the generation of edge or connection between the data points to form the neighborhood impacts the processing time of the algorithm. The complexity of this part of the modified algorithm is also challenged when the data points are not linear. The choice of value for r also is also critical as a minimum choice will produce more clusters which impact the inter-cluster separation.

B. Cluster Validation

As there is no optimal clustering algorithm [31], it is necessary to evaluate the generated clusters of the mining process on the International Data. One approach is an internal validation with which the concentration is the partitioned data such that the compactness and separation of the clusters are measured. The Silhouette index [32] is where the silhouettes show which objects lie well within their partition and which are somewhere between clusters. The silhouettes herein were formed basically by knowing the clusters or partitions generated by the modified clustering algorithm and the distance between the data points–hence, a data point i's distance to other points within the cluster it belongs to and to other data points in other clusters.

The average distance a(i) of a data point i to all other objects in the cluster it belongs to is computed in the same manner that the average distance b(i) to other objects in other clusters is also derived. Hence, the silhouette score is derived as:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)} \quad (3)$$

The Silhouette index is chosen for the validation of the resulting clusters of the proposed graph-theoretic clustering algorithm in order to observe how well the algorithm partitioned the data set [33]. The focus is also on the quality of the clustering structure being measured only using information or feature intrinsic to the data set [34]. Another salient point in choosing the Silhouette index for cluster validity is because it measures attributes taken from the data, itself and the clusters found [35]. The silhouette scores ranging from -1 <= s(i) <= 1 can be interpreted in Table III.

The validation on the clustering result generated by the modified graph-theoretic clustering algorithm infused with small-world network structure based on the Silhouette score is presented in Table IV which presents the silhouette score of the clustering result. The average intra-cluster distance was derived

from calculating the distance of a random point (x, y) in a cluster towards all other data points in the same cluster to which it belongs to. Inter-cluster distance is the distance of this (x, y) towards the other data points in other clusters.

TABLE III. SILHOUETTE SCORE INTERPRETATION

Range	***Description***	***Interpretation***
0.71 – 1.00	Strong	A strong structure has been found.
0.51 – 0.70	Reasonable	Reasonable structure has been found.
0.26 – 0.50	Weak	Structure is weak and could be artificial.
≤ 0.25	Not Substantial	No substantial structure has been found.

TABLE IV. SILHOUTTE VALIDATION ON CLUSTERING RESULT

Cluster ID	***Average Intra-Cluster Distance***	***Average Inter-Cluster Distance***	***Silhouette Score***
1	2.57	14.65	0.81
2	2.46	9.65	0.75
3	3.09	15.82	0.80
4	3.20	19.83	0.84
5	2.86	16.64	0.83
6	2.70	15.82	0.83
7	-	15.11	0.00

The average Silhouette score derived as 0.69 indicates an acceptable structure was found which is also manifested in the scores of all the clusters which derived scores not lower than 0.75 which means that each cluster has a strong structure except for Cluster 7 which has only one (1) data point – hence, silhouette score is 0. Such constraint is present to prevent the number of groups from significantly increasing [36]. Consequently, when a clustering result is interpreted based on Table III, the clustering is acceptable when the score is at least 0.50 [37].

IV. CONCLUSION

This work performed data mining in the international linkages of Philippine Higher Education Institutions (PHEIs) data using a proposed modified Prim's MST-based clustering algorithm producing a minimum spanning tree for the data set infusing the computation of local clustering coefficient for the data points in the limited neighborhood generated by von Neumann's Neighborhood.

An integral part of this work was the preparation of the raw data to achieve the dataset that is ready for processing by the modified Prim's MST-based clustering algorithm. The numerical attributes of the International Linkages dataset were used for the clustering to work on similarity on a particular parameter.

The results of the study show that there is an acceptable structure found in the clustering result with silhouette score 0.69 and 0.75 being the least score for the 6 out of 7 clusters derived for r=5 of the von Neumann Neighborhood.

However, the algorithm is still bound by the *a priori* input value of r which dictates the number of possible neighbors in one cluster for the von Neumann Neighborhood. As such the optimum number of clusters and most ideal value of r for a particular size of data are interesting.

Also for future works, the interest is also centered on the cluster validation utilizing external validity indices particularly those which works or are specific to graph-theoretic clustering algorithms. The data can also be refined more and subjected to clustering process to compare the performance of the traditional and the modified clustering algorithm.

ACKNOWLEDGMENT

The authors would like to extend gratitude to the Commission on Higher Education International Affairs Staff (Philippines) for cooperation in the realization of this work by providing the data necessary for the study. Appreciation is also extended to the Tarlac Agricultural University as well as the Technological Institute of the Philippines – Quezon City.

REFERENCES

[1] U. Teichler, "Internationalisation of higher education: European experiences," Asia Pacific Educ. Rev., vol. 10, no. 1, pp. 93–106, 2009.

[2] J. Lawrence, "Internationalization of Higher Education in the United States of America and Europe: A Historical, Comparative, and Conceptual Analysis (review)," Rev. High. Educ., vol. 27, no. 2, pp. 281–282, 2004.

[3] D. Dutschke, "Campus Internationalization Initiatives and Study Abroad," Coll. Forum, vol. 45, no. October, pp. 67–73, 2009.

[4] W. Green and C. Whitsed, "Critical perspectives on internationalising the curriculum in disciplines," Crit. Perspect. Int. Curric. Discip. Reflective Narrat. Accounts from Business, Educ. Heal., no. February, 2015.

[5] A. Aerden, F. De Decker, J. Divis, M. Frederiks, and H. de Wit, "Assessing the internationalisation of degree programmes: Experiences from a Dutch-Flemish pilot certifying internationalisation," Compare, vol. 43, no. 1, pp. 56–78, 2013.

[6] J. Knight, "Internationalization Remodeled: Definition, Approaches, and Rationales," J. Stud. Int. Educ., vol. 8, no. 1, pp. 5–31, 2004.

[7] V. Vijay, V. P. Raghunath, A. Singh, and S. N. Omkar, "Variance based moving k-means algorithm," Proc. - 7th IEEE Int. Adv. Comput. Conf. IACC 2017, no. i, pp. 841–847, 2017.

[8] B. Kenidra, M. Benmohammed, A. Beghriche, and Z. Benmounah, "A Partitional Approach for Genomic-Data Clustering Combined with K-Means Algorithm," 2016 IEEE Intl Conf. Comput. Sci. Eng. IEEE Intl Conf. Embed. Ubiquitous Comput. 15th Intl Symp. Distrib. Comput. Appl. Bus. Eng., pp. 114–121, 2016.

[9] J. Chang, J. Luo, J. Z. Huang, S. Feng, and J. Fan, "Minimum Spanning Tree Based Classification Model for Massive Data with MapReduce Implementation," 2010 IEEE Int. Conf. Data Min. Work., pp. 129–137, 2010.

[10] P.-N. Tan, M. Steinbach, and V. Kumar, "Chap 8 : Cluster Analysis: Basic Concepts and Algorithms," Introd. to Data Min., p. Chapter 8, 2005.

[11] R. Xu and D. Wunsh II, "Survey of clustering algorithms for MANET," IEEE Trans. Neural Networks, vol. 16, no. 3, pp. 645–678, 2005.

[12] C. Zhong, D. Miao, and R. Wang, "A graph-theoretical clustering method based on two rounds of minimum spanning trees," Pattern Recognit., vol. 43, no. 3, pp. 752–766, 2010.

[13] X. Wang, X. L. Wang, and J. Zhu, "A new fast minimum spanning tree-based clustering technique," IEEE Int. Conf. Data Min. Work. ICDMW, vol. 2015–Janua, no. January, pp. 1053–1060, 2015.

[14] L. Galluccio, O. Michel, P. Comon, M. Kliger, and A. O. Hero, "Clustering with a new distance measure based on a dual-rooted tree," Inf. Sci. (Ny)., vol. 251, pp. 96–113, 2013.

[15] G. W. Wang, C. X. Zhang, and J. Zhuang, "Clustering with Prim's sequential representation of minimum spanning tree," Appl. Math. Comput., vol. 247, pp. 521–534, 2014.

[16] P. Das and K. A. A. Nazeer, “A novel clustering method to identify cell types from single cell transcriptional profiles,” Procedia Comput. Sci., vol. 132, no. Iccids, pp. 983–992, 2018.

[17] D. R. Edla, S. Machavarapu, and P. K. Jana, “An Improved MST-based Clustering for Biological Data,” pp. 42–47, 2012.

[18] N. Paivinen, “Clustering with a minimum spanning tree of scale-free-like structure,” vol. 26, pp. 921–930, 2005.

[19] C. Zhong, D. Miao, and P. Fränti, “Minimum spanning tree based split-and-merge: A hierarchical clustering method,” Inf. Sci. (Ny)., vol. 181, no. 16, pp. 3397–3410, 2011.

[20] A. Singh, A. Yadav, and A. Rana, “K-means with Three different Distance Metrics,” Int. J. Comput. Appl., vol. 67, no. 10, pp. 13–17, 2013.

[21] C. B. Abhilash, K. Rohitaksha, and S. Biradar, “A Comparative Analysis of Data sets using Machine Learning Techniques,” Adv. Comput. Conf., pp. 24–29, 2014.

[22] D. Elsayad, A. Khalifa, M. E. Khalifa, and E. S. El-Horbaty, “An Improved Parallel Minimum Spanning Tree Based Clustering Algorithm for Microarrays Data Analysis,” no. Infos, pp. 66–72, 2012.

[23] L. H. Yen and Y. M. Cheng, “Clustering coefficient of wireless ad hoc networks and the quantity of hidden terminals,” IEEE Commun. Lett., vol. 9, no. 3, pp. 234–236, 2005.

[24] M. R. Brust, D. Turgut, C. H. C. Ribeiro, and M. Kaiser, “Is the clustering coefficient a measure for fault tolerance in wireless sensor networks?,” IEEE Int. Conf. Commun., pp. 183–187, 2012.

[25] C. H. Zeng and K. C. Chen, “Clustering coefficient analysis in large wireless ad hoc network,” 2017 IEEE/CIC Int. Conf. Commun. China, ICCC 2017, vol. 2018–Janua, no. Iccc, pp. 1–6, 2018.

[26] Q. K. Telesford, K. E. Joyce, S. Hayasaka, J. H. Burdette, and P. J. Laurienti, “The Ubiquity of Small-World Networks,” vol. 1, no. 5, 2011.

[27] D. A. Zaitsev, “A generalized neighborhood for cellular automata,” Theor. Comput. Sci., vol. 666, no. November, pp. 21–35, 2017.

[28] U. Sahin, S. Uguz, H. Akin, and I. Siap, “Three-state von Neumann cellular automata and pattern generation,” Appl. Math. Model., vol. 39, no. 7, pp. 2003–2024, 2015.

[29] N. Y. Soma and J. P. Melo, “On irreversibility of von Neumann additive cellular automata on grids,” Discret. Appl. Math., vol. 154, no. 5 SPEC. ISS., pp. 861–866, 2006.

[30] X. Min, X. Xu, and Z. Wang, “Combining von neumann neighborhood topology with approximate-mapping local search for ABC-based service composition,” Proc. - 2014 IEEE Int. Conf. Serv. Comput. SCC 2014, pp. 187–194, 2014.

[31] O. Arbelaitz, I. Gurrutxaga, and J. Muguerza, “An extensive comparative study of cluster validity indices,” vol. 46, pp. 243–256, 2013.

[32] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, no. C, pp. 53–65, 1987.

[33] J. Shen, S. I. Chang, E. S. Lee, Y. Deng, and S. J. Brown, “Determination of cluster number in clustering microarray data,” Appl. Math. Comput., vol. 169, no. 2, pp. 1172–1185, 2005.

[34] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. C. Tseng, “Evaluation and comparison of gene clustering methods in microarray analysis,” Bioinformatics, vol. 22, no. 19, pp. 2405–2412, 2006.

[35] D. N. Campo, G. Stegmayer, and D. H. Milone, “A new index for clustering validation with overlapped clusters,” vol. 64, pp. 549–556, 2016.

[36] F. Wang, J. D. Kelleher, J. Pugh, and R. Ross, “An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity,” 2017.

[37] R. C. De Amorim and C. Hennig, “Recovering the number of clusters in data sets with noise features using feature rescaling factors,” vol. 324, no. 2015, pp. 126–145, 2016.

Towards a Bespoke Document Tracking System for Philippine Higher Educational Institutions

Sheila R. Lingaya

Abstract*: This study on the development and validation of a document tracking model for utilization of Philippine Higher Education Institutions was undertaken to produce a system that would facilitate the management of documents in state universities or colleges by providing a way to monitor, record and track the location of in-process documents to support an academic organization. The Software Development Process was used as basis for the development of the software involving phases such as user requirements specification, design and implementation, validation and evolution (i.e. the process of changing or modifying the system once it has gone validation and yielded feedbacks for further modification). The acceptability of the software as evaluated by forty (40) office personnel representing every units of the Tarlac Agricultural University – the sample locale of the study, was confirmed in terms of user interface and functionality. These evaluators judged the software based on their skills and ability to use the software while carrying out their job functions. Five (5) IT experts also judged the software in terms of user interface, functionality, database design and security. Based on the results of the study, findings indicate that the document tracking system is excellent for the evaluators as process owners with a grand mean of 4.54 with its ease of use because of the simplicity of operations and the design itself with the reliability and usability or fitness for purpose as to tracking in-process documents and generating reports. The experts also evaluated the system as excellent with a grand mean of 4.58 – hence, the system's visual, functional and navigational elements and the manner it requests information helps the user operate the document tracker. Security was also judged as excellent because the system can control users and produce integral records.

***Index Terms*: document tracking, information system, transaction processing system**

I. INTRODUCTION

Bureaucracy is expected in most countries and most often than not, it had caused much effect in anyone's perspective of efficiency and effectiveness. In the Philippines, government agencies now are always under the watchful eye of the discerning public. They are always trying to become more efficient and effective in their delivery of services, primarily because they owe the public. The Republic Act 9465 or the Anti Red-tape Act has been issued which states that government transaction such as applications or renewal of permits, licenses and other documentation should be completed in five working days especially for simple cases and 10 working days for more complex transactions or requests.

Revised Manuscript Received on July 06, 2019.

Sheila R. Lingaya, Department of Computer Studies, Tarlac Agricultural University, Tarlac, Philippines.

It said that each agency is also required to reply to the client, whether requests are rejected or couldn't be processed with the explanations why it was rejected and what could be done to re-file their requests. Signatories in each document, the law states, must be limited to a maximum of five persons to reduce time and simplify procedures. As such, many perceive information and communications technology as a cost effective and convenient means to promote openness and comprehensive transparency efforts in most countries [1]. Information Systems are being employed and implemented to reflect improvement and efficiency as well as becoming at pace with the influx of technology. The trend nowadays is to include less paper and manpower in the organization's or institution's operation. Yet, although computer generated electronic records have been around, the phenomenon of a paperless office is still remote although paper consumption puts substantial pressure on today's world forest ecosystem where it seems on the face of it that emergence of computer and capacity of storage of documents in electronic form may lead to decrease in consuming such – hence, emergence of the paperless office [2]Apparently, any business organization or even an education institution still relies on standard operating procedures which primarily include pertinent documents and communications which do need to be managed efficiently and effectively in a manner that they can be tracked down or monitored. Even academic institutions such as the Tarlac Agricultural University boast their transparency of rendering services to their clientele, employees, office units and external community. It caters to the needs of its stakeholders via standard procedures which include or involve processing or pertinent documents. Management of 'in-process' documents would serve as a breakthrough in the manual operation of managing document's passing through the offices to minimize the problems encountered in following-up, tracking down and monitoring documents throughout the University – thus, a Document Tracking System. A document tracking assumes that knowing the movement of a document would enable a decision-maker to pinpoint where it is and in what state – thus, receiving, immediate feedback to make timely and rational decisions. It is a means for monitoring a document's movement from "birth" to "growth" to "death" and in some cases, to "rebirth." In relation to proper management process, this concept of record lifecycle sees records as passing through various stages: creation, active use, inactive use and then onward to either destruction or rebirth in the form of archives.

Another variation of document tracking is video based where the tracking of paper documents is on the desk over time and automatically linking them to the corresponding electronic documents using an overhead camera [3].

As such, DTS is a type of information system that handles the task of recording and monitoring in-process documents. Concerned with 'moving' documents, attributes of the document are captured into the system and not the document itself – profile of the object [4]. Having, this, a document tracking system when being developed needs to be in accordance with the type of business or organization or for this case, educational institution for which it is being developed for.

This is because of their uniqueness in business processes or the ways documents are handled or passed through.

The university observes manual taking note or records of documents and communications via "the logbook" monitoring or the "received by" and "released by" on a certain date system. Since computerized systems are necessities in an organization's way of accomplishing transactions and processes, the rate of adoption of electronic alternative over the past years and the dominance of paper over digitized records also justify the conduct of this work to facilitate the in-process documents' management of TAU as a Philippine Higher Education Institution.

This study aimed to develop and validate a Document Tracking System (DTS) for the University which will facilitate the movement of documents from one unit or office to another in the University and keep track of the whereabouts of these documents in process.

This paper is organized as follows. Section II presents the review of related literature. Section III focuses on the work's methodology followed by the presentation, analysis and interpretation of results of the study in Section IV. The Section V summarizes and concludes the paper and some future works.

II. METHODOLOGY

A document is an identifiable recording of information and any recording medium can be used as long as it persists over time. Information is more than the data. So a document includes some elements of contextualization, organization and analysis and even if one's job is just the management of documents for some specific corporate purpose, it is a professional responsibility to know the relationship of those documents to the society [5].

The most important factor for the success of this project was how closely the particular plan was defined and followed. In order to at least be as close to achieving such, this study was defined with a schedule to follow for its development from its birth and eventually to full development, towards the in-depth analysis of the possible processes that it could offer as features to solve the problems encountered in the current 'in-process' document management of the University. The development was guided by the Concept of Software Development Process with fundamental activities, namely: Specification, Design and Implementation, Validation and finally, Evolution. The "evolution" in this study was the idea of correcting the errors based on feedbacks of the validation phase.

A. Data Gathering Procedure

In order to analyze the performance of the proposed Document Tracking System, there is a need for appropriate materials or instruments to collect pertinent data. Observations, interviews and questionnaires are the most appropriate for this purpose in this study. During this study, the observation took charge on investigating the available facts and data to obtain specific objectives. The researcher eyed the process or tasks involved in university's document management.

Interviews were also employed to facilitate the acquiring of the pertinent and supplementary data that may not have been gathered during the observation. These data primarily were specific on the parts of a Document Tracking System namely: the people (operator, management), equipment (computer, printer, barcode readers), data (from the documents), tools, space (office units), and procedures. The questionnaire was used to gather information and opinions from the end-users of the proposed-systems. The respondents of this questionnaire were given a background of the proposed system or actually were allowed to experience the proposed system's design.

B. Development Tools

The system was developed utilizing MySQL, an open source Database server and a relational database management system that works in client/server or embedded systems. MySQL, the most popular Open Source SQL database management system, is developed, distributed, and supported by MySQLAB. The choice was for its main features of including portability, security, and scalability. PHP was used as the Web Development language - a server-side scripting language, which can be embedded in HTML or used as a standalone binary. This scripting language includes features such as it is free, easy to use, HTML-embedded, none-tag based, stability, speed, extension to other programs and protocols, fast feature development, popularity and non-proprietary. With these, the system was made more stable as it was developed as a web-based system.

C. Software Validation

To determine the efficiency of the developed software, the following scale was used by the IT experts and users in rating the system. To determine the adequacy of scope and user-friendliness of the developed software, the users used the scale presented in Table 1. This ensured that the true requirements of the system were yielded by exposing it to potential end-users.

Table 1. Scale used in Evaluating the System

Units of Indexes	Adjective Description
4.50 – 5.00	Excellent
3.50 – 4.49	Very Satisfactory
2.50 – 3.49	Satisfactory
1.50 – 2.49	Poor
0 – 1.49	Very Poor

III. RESULTS AND DISCUSSION

A. The Existing System

The flow of operations of the existing system of managing in-process document in the university is depicted in the Figure 1. It clearly depicts that some activities in the process of recording the in-process documents are repetitive in a way that clerks who are in–charged as receiving and releasing officers repetitively record or note about their incoming and outgoing documents. The dataflow diagram of the existing system in Figure 1 indicates that documents do not have in any way a unique identification of itself in the process which may be used in tracking it. This is because the same single document can be recorded differently by the different releasing and receiving officers with the aforementioned procedure.

There is no definite way to track or search the document or worse, to know the document's location. To hunt the document through the log books is a complex process since one would have to look into receiving and/or releasing logbooks of every office where the document may have passed. Added to the burden would be on searching more specifically for a document on the notes recorded on the logbook. The process boils down to looking onto logbooks after logbooks and records after records.

Figure 1. Dataflow Diagram of Existing Document Management of In-Process Documents

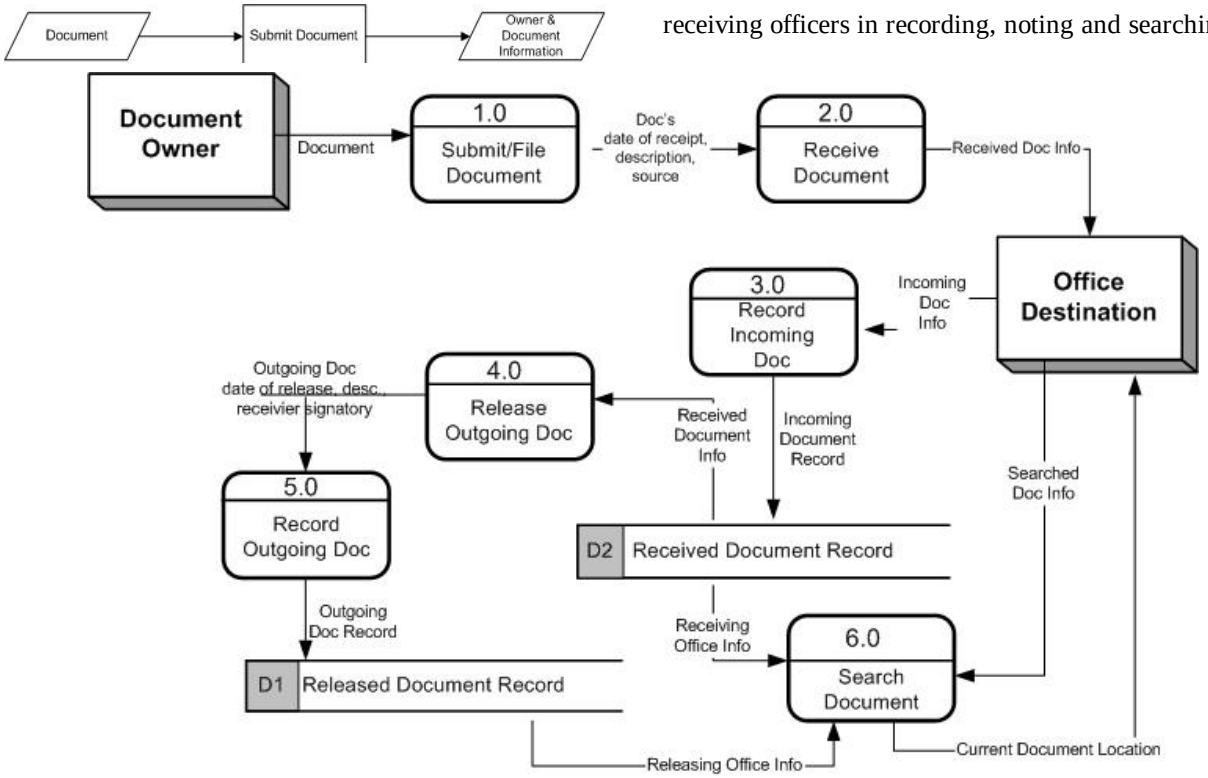

Figure 2. Input-Process-Output of Document Tracking

The releasing and receiving officers in an office in the units are not always the clerk. The existing process does not have means to record who may have released or received the document so that when time comes that a document being searched or tracked down is identified to be last received in an office, another question would be who received it.

B. Design and Implementation

After careful analysis of the existing system based on gathered facts, the researcher was able to develop the following one or more different system models and prototypes to depict the proposed system's flow of operations.

Figure 2 shows the process involved in the proposed system which actually depicts the flow of operations of the document tracking process using the proposed system.

The creation of an office account and generation of barcode is the definite difference of the existing system and the proposed Document Tracking System.

The process of generating a barcode plays a significant role in the tracking or searching of the document. It is the unique identity to be used by the document owner, releasing and receiving officers in recording, noting and searching for the document's current location.

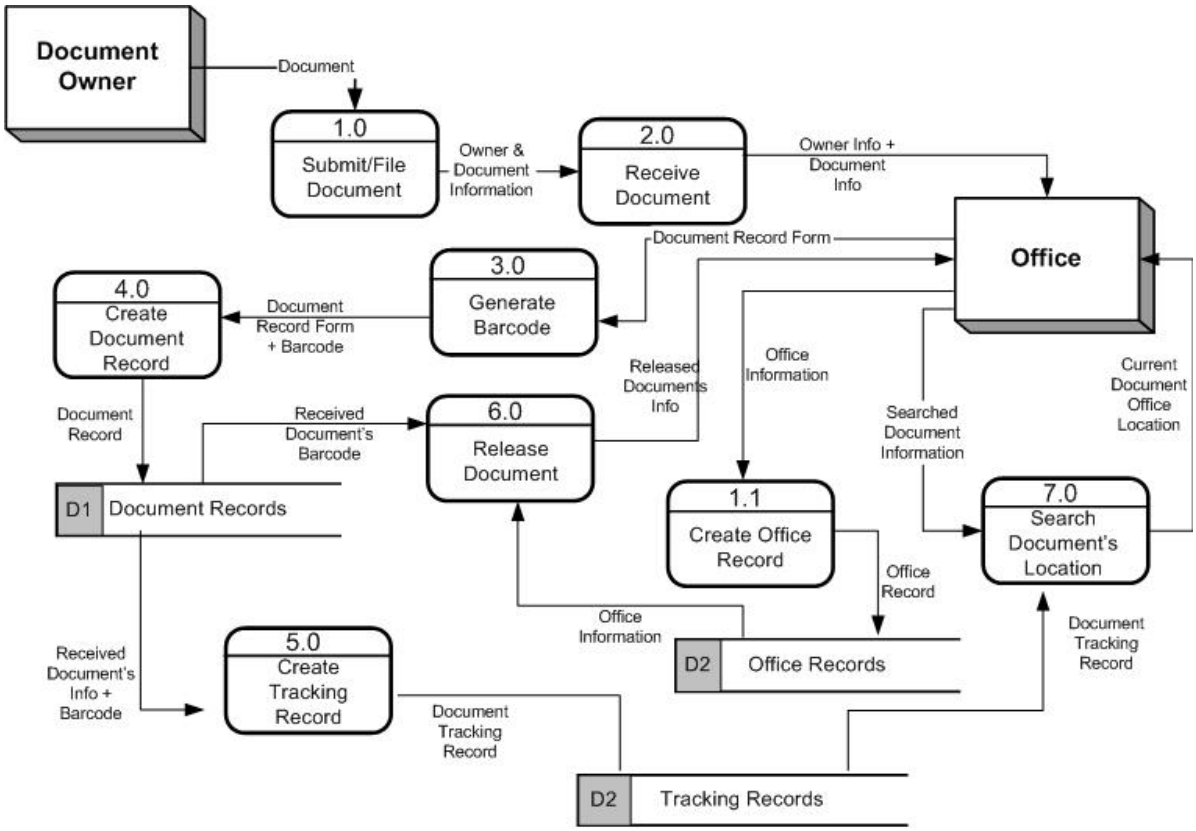

Figure 3. Dataflow Diagram of Document Tracking System for In-Process Documents

Once a document is received by an office, the document's record is created to note its owner and to generate a barcode image which will serve as its identification. This process only happens once for a document and only at the first receiving office. Using the document's id, the receiving office shall then be tagged – using the barcode reader to recognize the printed barcode id – to signify that the document was received by the office. This replaces the process of having to incoming document over and over again from one office to another. The barcode id shall also be used to tag the release of the document by the office.

The dataflow diagram in Figure 3 depicts the flow of data from one process towards the next. It also emphasizes the destination and origin of services and information during the document tracking process. As an information system being defined by a well-organized database structure that will

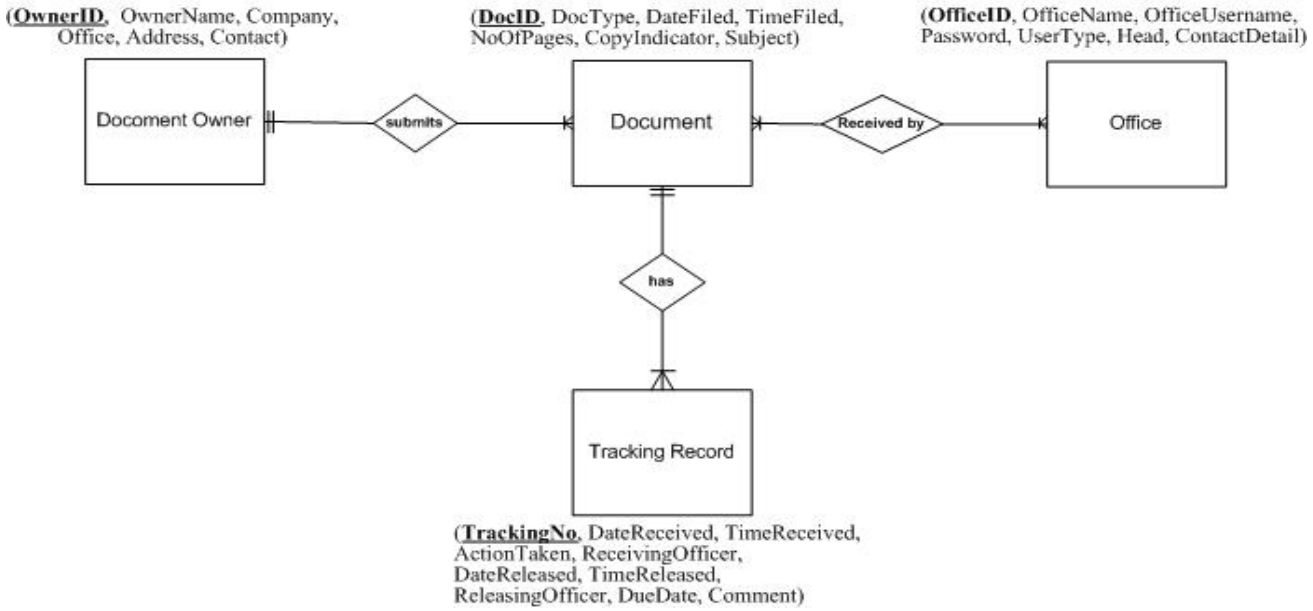

Figure 4. Entity Relationship Diagram of the Document Tracking System.

record an

handle the entry of data into the system and how it will generate necessary reports, the entity relationship diagram in Figure 3 displays how the back-end of the proposed system was designed to hold the information that shall enter the system.

The database was designed to be centralized mainly for the database's maintainability. Since the document tracking system emphasizes on how timely records or information are synchronized for them to be available to user, the database was designed to a centralized database as back-end. Figure 4 displays the Entity Relationship Diagram of the Document Tracking System.

The synchronization of records or information is critical in the tracking feature of the online document tracking system since an update in records in one office should be immediately recognized by another by any chance of dealing with same document that is being processed or reports being produced.

A tracking record is created once a document is received by an office either from another office or from the owner himself. The tracking records table is the main table of the database for it is from which that the current location or the last receiver can be identified. The fields with which the tracking record contains no DateReleased, TimeReleased and ReleasingOfficer clearly indicates that the office to which this document's tracking records was created is the document's current location.

Figure 5. Barcode of a Document

As seen in Figure 5, a user can print a document's barcode ID in the physical document itself or on a sticker and recognize it also to produce reports out of said information primarily to track the document itself.

C. Validation Results

Five Information Technology experts of proven expertise in web applications development and management were tapped to judge the Document Tracking System. Their comments and suggestions were considered in the improvement of the system.

The experts were composed of one net administrator, two web developers and one systems administrator from government institutions namely: Central Luzon State University (CLSU), and the target itself, TAU. One web developer coming from the Manila-based Global Property Guide, a private company, also evaluated the system. All of them are highly qualified in evaluating the system because of their familiarity with web applications and administration. The questionnaire for the evaluation was based on the criteria specified in the statement of objective of this study. Table 2 indicates the experts' evaluation on the user interface.

Table 2. Experts Evaluation of User Interface

Software Evaluation Criteria		Average	Descriptive Rating
A. User Interface **(Composite Mean : 4.52)**			
A.1	Visual Appearance	**4.40**	Very Satisfactory
A.2	Appropriateness of design	**4.40**	Very Satisfactory
A.3	Navigational elements	**4.60**	Excellent
A.4	Request for information	**4.60**	Excellent
A.5	Functionality of barcode reader	**4.60**	Excellent

The IT experts came up with the aforementioned rating by assessing the Document Tracking Systems interfaces, online. The system was accessed on http://doctrack.tca.edu.ph on a terminal owned by the IT expert evaluator as assisted by the researcher or the web administrator of the College.

Table 3. Experts Evaluation of Database Design

Software Evaluation Criteria		Average	Descriptive Rating
B. Database Design **(Composite Mean : 4.60)**			
B. 1	Arrangement of data	**4.60**	Excellent
B. 2	Synchronization of database	**4.60**	Excellent
B. 3	Logical Design	**4.60**	Excellent

The tracking records generated by the Track Document feature of the system indicate how the data were arranged in the system and how they are presented as evaluated by the IT experts with which result of evaluation is indicated in Table 3.

Table 4. Experts Evaluation of Security

Software Evaluation Criteria		Average	Descriptive Rating
C. Security **(Composite Mean : 4.60)**			
C. 1	Authorization of User	**4.60**	Excellent
C. 2	Implementation of user permissions	**4.60**	Excellent
C. 3	Integrity of records	**4.60**	Excellent

An indication of security noted by the IT experts as indicated in Table 4, during the evaluation was the difference on the permissions of a regular office account and that of an administrator account.

The results of the IT expert's evaluation came up with an over-all mean of 4.58, interpreted *Excellent* based on the scale used. It consists of the composite means of 4.52 for user interface, 4.60 for functionality, 4.60 for database design and 4.60 for security. Meanwhile, the researcher submitted the system for evaluation to end-users such as clerks, office personnel, as well as

on-the-job trainees and student assistants in the office/units since they also take charge in recording the incoming and outgoing documents of their offices. One representative from the forty offices in the university was tapped to evaluate the system purposively including students who also submit reports/documents to the offices as part of student organizations such as student publications and student councils.

The result on their evaluation on the ease of using the system is indicated in Table 5.

Table 5. Users Evaluation on Ease of Use

Software Evaluation Criteria		Average	Descriptive Rating
A. Ease of Use **(Composite Mean: 4.54)**			
A.1	Simplicity of Design	4.55	Excellent
A.2	Simplicity of Operation	4.53	Excellent

Some of user evaluators also gave their comments which were considered by the researcher in further improving the design in order to meet the end-users' requirement to the system. The Record's Office head was made to evaluate the system using the Administrator account while the rest of the offices were given their own individual office accounts with username and password. The Document Tracking System was uploaded online for this purpose. The users were provided their own accounts per office and oriented on how to use the system. The questionnaires were handed on them and was made to reply on their own initiative on whether their experience of the system is already enough to give it their ratings. Table 6 also shows the users evaluation on Usability.

Table 6. Users Evaluation on Usability

Software Evaluation Criteria		Average	Descriptive Rating
B. Usability (**Composite Mean: 4.58**)			
B.1	Keeping track of incoming documents	4.70	Excellent
B.2	Keeping track of outgoing documents	4.65	Excellent
B.3	Tracking in-process documents	4.65	Excellent
B.4	Generating reports	4.65	Excellent

The use of the barcode scanner to read the barcode document ID was demonstrated to the user evaluator when tracking the document, receiving the document and also in releasing the document. Users gave positive feedback on said process because it eliminated the tedious repetition of recording the incoming and outgoing documents from office to office and from logbooks to logbooks.

The *receive document* feature of the Document Tracking System affirms the elimination of repeatedly recording received documents in all offices. The idea of a document being recorded as received on all offices and traveled is replaced by the action of just tagging the document through scanning the barcode document ID.

Table 7. Users' Evaluation on Reliability

Software Evaluation Criteria		Average	Descriptive Rating
C. Reliability (**Composite Mean: 4.51**)			
C.1	Response to user actions	4.48	Very Satisfactory
C.2	Message Prompts	4.48	Very Satisfactory
C.3	Reports Generation	4.58	Excellent
C.4	Management/presentation of outputs or reports	4.50	Excellent

Reliability of the system as evaluated by the users which results are indicated in Table 7 and the IT experts connotes similar idea. It defines whether the users are being presented with reports that can be used officially by the system.

Reliability was also interpreted by the users whether it prompts messages or gave appropriate response to their actions as well as how the outputs or reports are presented. The results of the users' evaluation were a composite mean of 4.54 for ease of use, 4.58 for usability, and 4.51 for reliability. Over-all, the system obtained a mean of 4.54.

IV. CONCLUSION

The Document Tracking System developed as a bespoke model for Philippine State Universities and Colleges was developed based on the Software Development Process activities. The IT experts gave a positive feedback on the user interface, functionality, database design and security with a grand mean of 4.58 or excellent. The acceptability of the software was evaluated by end-users as to ease of use, usability and reliability and a positive response based on the aforementioned criteria was derived with a grand mean of 4.54 – hence, excellent. The range of values of the results of evaluation of the system shows that the evaluators judged the system highly acceptable in the provisions for enough dry run, in the statement of the desired validation outcomes, and in the usefulness and performance of Document Tracking System. This turns out to agree with the result of the studies cited in the related studies, specifically [4] which highlighted the underlying factors in the development and employment of a document tracking system which helped out in the development of the Document Tracking System.

For future work, the document records will be studied for purpose of data mining to optimize the features of tracking in-process documents. Also, the system is planned to be integrated on the University-wide document management system which also needs to be integrated with the information systems which depend on document and information retrieval.

REFERENCES

1. J. C. Bertot, P. T. Jaeger, and J. M. Grimes, "Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies," *Gov. Inf. Q.*, vol. 27, no. 3, pp. 264–271, 2010.
2. R. York, "Ecological paradoxes: William Stanley Jevons and the Paperless Office," *Hum. Ecol. Rev.*, vol. 13, no. 2, pp. 143–147, 2006.
3. J. Kim, S. M. Seitz, and M. Agrawala, "Video-based document tracking: unifying your physical and electronic desktops," *Proc. 17th Annu. ACM Symp. User interface Softw. Technol. (UIST '04)*, vol. 6, no. 2, pp. 99–107, 2004.
4. Loyola, *Toward a Common Computer-Based Document Tracking System in U.P. Diliman*. 2001
5. M. Gilheany, *Archive Planning, Analysis Newsletter for Document Management*, 2001

AUTHORS PROFILE

Sheila R. Lingaya is currently an Assistant Professor at the Tarlac Agricultural University – Tarlac, Philippines where she also serves as Assistant Director of External Linkages and International Affairs. She is also a student of the Doctor of Information Technology program of the Technological Institute of the Philippines – Quezon City. She also has to her name, a Master of Science in Information Technology from the Tarlac State University – Tarlac City, Philippines in 2011. Her research interest includes information systems, data mining and pattern recognition.