Decal : A Data Mining Clustering Technique with Report Customization

Michael Jessie Theodore A. Sese, Anna Liza A. Ramos, Archie A. Alonte, Ellen Christine A. Correa, Donna Faye H. Dilla, Lois Lyn B. Tompong, Camille F. Alzona

[1]Saint Michael's College of Laguna, Biñan Laguna, Philippines.
Email: mec@smcl.edu.ph; msese@smcl.edu.ph

ABSTRACT

The continuous growth of data per individual will soon become a problem. Technology makes our everyday living easier and faster, and also lead to the continuous growth of data but also gave the solution to it. Data Mining is then born. It is the process of analyzing data from different perspectives and summarizing it into useful information. It has great potential to help organizations to predict future trends and behaviors through patterns. Furthermore, clustering technique, which is under Data Mining, is used. Clustering is the grouping of a particular set of objects based on their characteristics, and combining them according to their similarities. It clusters to discover and minimize data into a new categories and groups. The need to extract useful information and to interpret extracted data is the two main reasons to use data mining as the demands for data increases. This study is to develop an online data mining clustering technique with report customization system that will enhance the data uploading, sorting, and duplicate-elimination, and report generation. The study will be beneficial for institutions to decrease time consumed and to increase productivity and satisfaction. This study will be able to solve the following problems: (a) the difficulty in accessing data from various departments; (b) the redundancy of data; and (c) the complexity of generating reports. The Researchers will use Agile Software Development (ASD) method to determine the needs on which the system will revolve and resolve, and will help the researchers to develop the features of the system.

Keywords : *Data, Information, Reports, Data Mining, Clustering*

1 INTRODUCTION

Students became focused to the fast service response which plays an important role in general industry today "[14], [16]". As time is a quantity that is non-renewable [9], any process that saves it is considered vital in many applications [1]. Technology made processing of records easy through computers [5]. Schools and universities would use modern technologies as a tool for improvement [19], like online systems which solves the problems regarding the processing of any records and satisfies the people connected in the institution [3][10]. The advancement of information technology enables organization to have a faster pace in the information exchange and productivity [7]. With the rapidly increasing volume of data, storing and retrieving data gets harder, thus, more automatic and effective mining approaches are required "[22],[13]". Nowadays, numerical data collected from databases and different data mining techniques have evolved and various algorithms evolved to mine the non-numerical data and the large volumes of heterogeneous data [15].

The most common problem in educational industries is the lack of data management systems that would minimize the time consumed on searching and retrieving needed data of different students through the years [6]. To resolve the problem, data mining technique was born [23]. It is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [29]. It is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis [24]. It is a powerful tool that can help you find patterns within your data and hidden information throughout your data sets "[28], [18]".

The main task of data mining is clustering [20]. It starts with the single observation clusters and progressively combines pairs of those which form smaller numbers of clusters that contain more observations and then clusters successively merged until the desired cluster structure is obtained [22]. As a fundamental task in Data Mining, its goal is to discover a new set of categories, the new groups are of interest in themselves and their assessment is intrinsic [26].

Project Context

The society had a lot of changes with the enrichment of technology. There is always a new process or innovation to replace one after the other. One of the benefactors of technology's great contributions to the world is to the educational system. Every organization uses manual system up until computers existed. The endless enrichment of technology gave birth to the Internet resulting to larger set of information. Ever since its birth, everything computerized connects online. The most common problem in educational industries is the lack of data management systems that would minimize the time consumed on searching and retrieving needed date of different students through the years. Thus, to resolve the problem, it resulted to the birth of data mining technique. Data mining is conceived primarily by research-driven tools focused on single task. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. It is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. It has great potential to help industries. It is a powerful tool that can help you find patterns and relationships within your data. Data mining discovers hidden information in your data, but it cannot tell you the value of the information to your organization. Data Mining Techniques shows how to quickly and easily tap the gold mine of business solutions lying dormant in information systems. Data mining tools predict future trends and behaviors, allowing industries to make proactive, knowledge-driven decisions. Data mining tools can answer questions that traditionally were too time-consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. The extraction of information from large data sets and transform it into understandable structure for further uses is the overall goal of data mining process. Understanding our world requires conceptualizing the similarities and differences between the entities that compose it. Thus, clustering is then born. Clustering is the grouping of a particular set of objects based on their characteristics, and combining them according to their similarities. It is a fundamental task in Data Mining. The goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. The core components of data mining technology have been under development for decades. The need to extract useful information and to interpret extracted data is the two main reasons to use data mining as the demands for data rapidly increases.

General Problem

The researchers found the need for a web-based data mining system that would ease the searching and retrieving of data. Long waiting lists of services in the queue have become a symbol of the inefficiency in terms of services.

Specific Problems

Difficulty in accessing data from various departments.

With eighty-six and sixty-seven hundredths percent (86.67%) of respondents said that the transferring of data from department to department is difficult. From comments stated that there is a poor communication between departments sometimes, and when mostly needed, which results to misunderstandings and longer processing time. Furthermore, through the observations, gathering those data for accessing is already a burden. Browsing through a large data set would be difficult, in result, it consumes time. Due to the increase in the amount of recorded data, it has become difficult to find, analyze patterns, associations within such large data [17]. Officials find it hard to deliver the desired service as fast as they could for retrieving the specific data [4].

Redundancy of Data.

With a one hundred percent (100.00%) of respondents who confessed of having a large set of data and having trouble in terms of avoiding and/or eliminating data duplicates in their records. While eighty-six and sixty-seven hundredths percent (86.67%) do not have an efficient system in terms of searching needed data. Furthermore, according to the comments, collecting data from different department causes duplication of data for a specific person, instead of fastening the process, it causes confusion and conflicts. The preprocessing of data is the initial and often crucial step of the data mining process [11]. Duplication of data shows inefficiency in data storing.

Complexity of generating reports.

With an eighty-six and sixty-seven hundredths percent (86.67%) of the respondents shared that their system is inefficient in terms of reports generation. Furthermore, according to the comments, creating necessary reports for the data gathered is “complex”. Reports are essential to portray and evaluate something. They are to further analyze the current status of the performance of the system [27].

Research Objectives

General Objectives

The researchers aim to develop an online data mining system that will increase the flexibility of information gathered and consolidate the records in different resources to avoid duplication of reports.

Specific Objectives

To consolidate data from various departments to make retrieving of data easier. The respondents see the system as an enhancement of the collection of data with 4.93 WM and look forward to the system which would ease the gathering of data with a 4.87 WM. Poor communication is a hindrance in the productivity of a department in an organization [26]. Therefore, the researchers will develop a system which can lessen the time in collecting and finding records. With a student information module, it is easier to collect, locate and view student records.

To eliminate redundancy of data. The respondents see an online system that could ease the duplication of data with 5.00 WM. As data is a vast collection of characters, numbers and special symbols, how could someone differentiate a data from another? Clustering was then recognized to find uniqueness to eliminate duplication of records [8]. Therefore, the researchers will develop a system that can merge the data from different resources with the same student number to avoid the redundancy of records.

To fasten the creation of reports. The respondents see an online system that would help them create reports efficiently with 4.60 WM and enhance their reports with 4.60 WM. Reports promote understanding, and faster compliance to any demand is good for any organization for it shows how good the system of your organization has [2][12]. Therefore, the researchers will develop a system that can create and customize reports to properly comply with the need and requirements of the user.

Significance of the Study The study will be a great help to the following:

To the Institution. The study will improve the record services thus provides proper management of reports, ensures accuracy and time efficiency in accessing the reports.

To the Officials. The study will be beneficial to the officials of the institution for it can enhance their response for each service and will help in reducing the workload. It also makes the reports more manageable and easier to control.

To the Students. The study will be beneficial to the students for it will fasten their request of services. Their time will be minimized.

To the Researchers. The study will be beneficial to the researchers as a guide to help them prepare for the related or further projects and researchers in their field.

To the Future Researchers. The study will be beneficial to the future researchers who may look into the possibility of developing a more improved and efficient system. This study can also serve as a reference that will help them gain basic knowledge for their future project.

Scope and Limitation

Scope The proposed system has only one end-user, the Admin. Admin (Administrator) is capable of accessing the system and the whole function of the system which are viewing, uploading, managing and accessing student records, and creating reports. In a sense, the admin had the full power to access the said system. Through the research done, the researchers had come up to develop the proposed system which would include the following features:

Students' Information Module - where the admin can view the records of the students. It has the main information of each student: Student ID No., First Name, Last Name, Middle Name, Age, Sex, Year, Course, etc.

Data Analysis Module - where admin can view the data which are represented by graphs. Data graphs can be adjusted by the quantity of imported data per year.

Import Data Module - where the admin can import Excel (.xls) files in the system from different resources/departments. It has

the basic information and specific information from its different resources. The subject, grades, units, borrowed status, charges, etc.

Generate Report Module - where the admin can generate the desired student report. It is divided into two modules:

Student Report - where the admin can select specific student, generate and print their records.

Customization - where the admin can drag and drop selective data column to customize specific information of students.

Aside from the mentioned features, which are used mainly for handling the data mining, other features are also included such as:

Log In Module - used to provide access to authorized users (admin) in order to maintain the security of the records and the entire system.

User Profile Module - consisted of the following:

View Profile - where the admin can view his/her profile information.

Change Username & Password - where the admin can modify the username and password of the administrator account.

Limitation

The proposed system is developed to provide a data mining system for institutions, schools, universities and colleges. It is a web-based application. It will only focus on data mining where agglomerative hierarchical clustering technique will be applied and as well as customization of reports. It gets all the data from all the departments in the institution and does not generate its own data. The system can extract Excel (.xls) files only. The Admin is the only authorized user of the system who can access and manipulate most of the features and usage of the system.

2. Methodology

2.1 Software Development Process Model

The researchers adapted the concept of Agile Software Development (ASD). ASD focuses on keeping code simple, testing often, and delivering functional bits of the application as soon as they're ready. The goal of ASD is to build upon small client-approved parts as the project progresses, as opposed to delivering one large application at the end of the project.

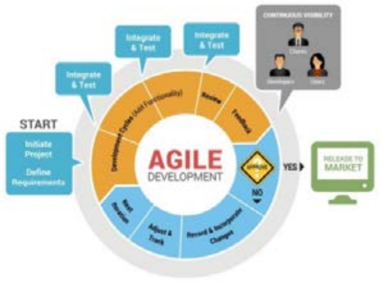


Figure 1. Agile Development Model

Planning Phase. The researchers identified the requirements needed by observing the situations and processes about the transaction and communication between departments. Different tasks were assigned in each member of the team in able to focus in each feature that will be needed in the system and iteration.

Development Phase. The researchers create, test, and demonstrate features.

Adapting Phase. The researchers review the output of iteration and re-plan based on discoveries. The researchers review the result and feedback in the system and refine features based on the feedback in iteration.

Deployment Phase. The researchers deliver the code to production environment, with all support needs in place. If the feedback on the system is satisfied or it is approved then it can now be released for bigger number of users. If not, the iteration should apply again

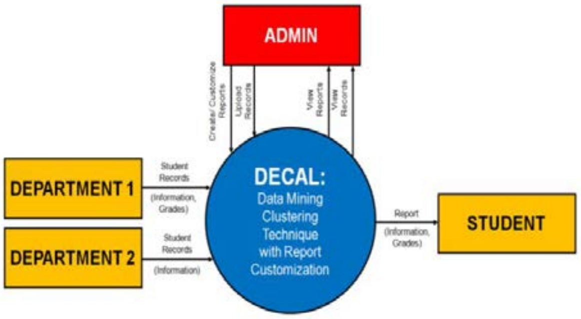


Figure 2. Decal Context Diagram

The figure shows the Context Diagram of the system. The diagram shows the only user of the system (Admin), along with the two sample departments and the two other intended benefactors of the system. The data, such as the student records and grades, will come from the departments of the school. The Admin will upload the records into the system and create reports from those records. The students of the school are the benefactors of the system, besides the admin. Through the help of the system, they could retrieve records and reports easily.

3 Results

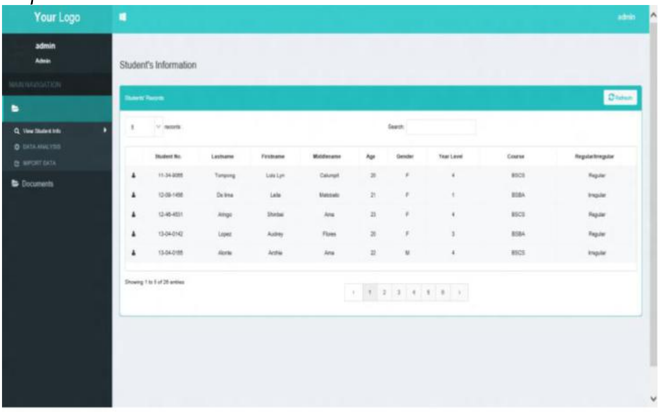
This section shows if the objectives of the study had been met based on the different tests, and survey conducted to assess the system.

Table 1. Software Evaluation Results of DECAL

All Characteristics	Mean	Verbal Interpretation
Functionality	4.83	*Excellent*
Reliability	4.71	*Excellent*
Usability	4.92	*Excellent*
Efficiency	4.92	*Excellent*
Portability	5.00	*Excellent*
Compliance	5.00	*Excellent*
Total Weighted Mean	4.90	*Excellent*

In evaluating the Functionality, it earned an average mean of 4.83 or interpreted as Excellent. Reliability earned an average mean of 4.71 or interpreted as Excellent. Usability earned an average mean of 4.92 or interpreted as Excellent. Efficiency earned an average mean of 4.92 or interpreted as Excellent. Portability earned an average mean of 5.00 or interpreted as Excellent. Compliance earned an average mean of 5.00 or interpreted as Excellent. In general, all characteristics earned an average mean of 4.90 or interpreted as Excellent.

To consolidate data from various departments to make retrieving of reports easier.



Student's Information

Students' Records

5 records

	Student No.	Lastname	Firstname	Middlename	Age	Gender	Year Level	Course	Regular/Irregular
♟	11-34-0085	Tompong	Lois Lyn	Calumpit	20	F	4	BSCS	Regular
♟	12-09-1498	De lima	Lala	Matsalo	21	F	1	BSBA	Irregular
♟	12-46-4531	Aringo	Shirlee	Ana	23	F	4	BSCS	Regular
♟	13-04-0142	Lopez	Audrey	Flores	20	F	3	BSBA	Regular
♟	13-04-0185	Alonte	Archie	Ana	22	M	4	BSCS	Irregular

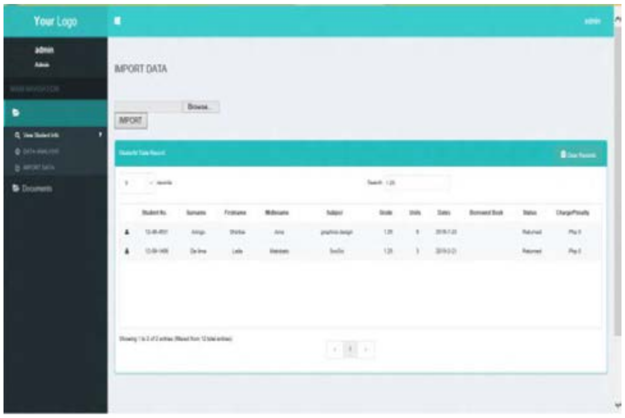
Showing 1 to 5 of 26 entries

‹ 1 2 3 4 5 6 ›

Figure 3. Student Information Module

The Student Information Module contains the basic information of the student. It provides search box to easiy find student information

To eliminate the redundancy of data.



IMPORT DATA

Students' Data Record

5 records

Search: 1.25

	Student No.	Surname	Firstname	Middlename	Subject	Grade	Units	Dates	Borrowed Book	Status	Charge/Penalty
♟	12-46-4531	Aringo	Shirlee	Ana	graphics design	1.25	6	[illegible]		Returned	Php 0
♟	12-09-1498	De lima	Lala	Matsalo	[illegible]	1.25	3	[illegible]		Returned	Php 0

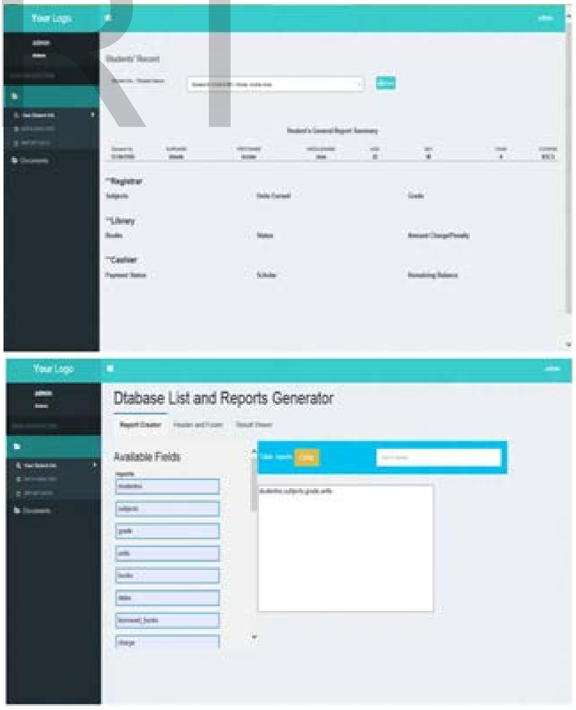
Showing 1 to 2 of 2 entries (filtered from 12 total entries)

‹ 1 ›

Figure 4. Import Data Module

This module allows the user to import excel files directly to the system. The system will analyze the data and made a clustering technique particularly information that pertains to a particular record. In addition, the system will automatically distinguish the Student ID number that serves as unique primary that determine that a particular record belongs to an individual assigned.

To fasten the creation of reports.



Students' Record

Student's General Report Summary

*Registrar

Subjects Units Earned Grade

*Library

Books Status Amount Charge/Penalty

*Cashier

Payment Status Scholar Remaining Balance

Dtabase List and Reports Generator

Available Fields

reports

studentno

subjects

grade

units

books

status

borrowed_books

charge

studentno,subjects,grade,units

Figure 5. Report Generator Module

The Report Generator Module allows the admin to generate specific report based on the preferences of the users because the systems allows you to select a specific field to be reflected in the report because it provides a customization functionality means the users can have the full authority to arrange the information accordingly.

4 Conlusions & Recommendation

4.1 Conclusions

The developed system assisted the admin to retrieve records faster at an efficient manner. The target user can extract reports in Excel (.xls) file format. He/she can print a customized report where they can select what report will be seen or printed out. Furthermore, the following were drawn: *To consolidate data from various department to make retrieving of data easier.* With the development of DECAL: Data Mining Clustering Technique with Report Customization System, the collection and retrieval of data from different department become easier. Through the use of the system uploading of records from different sources is faster and easir, same with searching and locating specific records.

To eliminate redundancy of data. The uploading of records doesn't cause any duplication of records anymore. By the use of the Clustering technique, records with similar Student ID number were clustered to combine records of a single records. *To fasten the creation of reports.* the creating and customizing reports is easier and faster yet efficiently supplied with the needed information. Through the system, the admin may either choose to create a report or customize a report. By creating a report, selecting a specific student name or ID is needed and by customizing report, select specific fields to show in your report, add some header and footer and print.

4.2 Recommendation

The study may contribute in an increased opportunity for further collaboration, and give direction for future studies. The study may also contribute to the general knowledge based information.

The researchers recommend this study the educational institutions with large quantity of data. It will help them to enhance their standards as far as document storage is concerned. Also, this research is recommended for future researchers who have interest in improving and developing system this particular system. Furthermore, this study will be a guide for them.

Acknowledgment

The researchers would like to sincerely thank Mr. Adrian Evanculla, Ms. Karla Mirasol Maranan for the Technical aspect of this paper and Ms. Inecita R. Cuevas for editing the manuscript.

References

[1] Adedayo, A. 2012. *Queuing Network Analysis of Patient Flow and Resource Allocation*. https://www.academia.edu/

[2] Dr. Ahmel, K. 2011. *Organizations and Technology*. International Journal of Advanced Computer Science.

[3] Anquer, T.U. 2012. *Advantages of Technology to the World*. https://www.scribd.com/

[4] Baricanosa, A. 2011. *Services and their Effect*. University of Perpetual Help System - JONELTA.

[5] Burgos, J. 2012. *Innovation in Today's Life*. https://journals.ateneo.edu/

[6] Dela Reyes, R.E. 2013. *The Efficiency of Data Mining*. https://josemariasison.org/

[7] Enrispe, L. 2016. *Benefits of Data Mining to an Organization*. https://infoplease.com

[8] Fazle, W.P. 2012. *MINIMIZING through CLUSTERING*. San Diego State University.

[9] Fuler, S.M. 2013. *Effective Mining Approaches*. University of Perpetual Help System - JONELTA.

[10] Haian, C. 2014. *Information Technology*. University of Cagayan East.

[11] Jztrik, G. 2015. *The Organization: Guide to a Better and Stronger Business*. https://www.academia.edu/

[12] Liebert, H. 2015. *A Good Organization*. https://www.academia.edu/

[13] Lior, S. 2013. *Efficiency of Using Data Mining to Respond to the Growing Data of an Organization*. https://www.scribd.com/

[14] Matag, M.N. 2013. *Fast Service for Fast Satisfaction*. https://www.scribd.com/

[15] Osmar, R.Z. 2016. *Principles of Knowledge Discovery in Databases*. https://webdocs.cs.ualberta.ca/

[16] Pangilinan, M.D. 2010. *General Industry*. https://prezi.com/

[17] Pascual, B. and Reaso, C. 2015. *The Increase of Data*. University of Perpetual Help System - JONELTA.

[18] Russell, D. 2015. *Data Mining*. www.mrlocke.com/

[19] Sarmiento, A.B. 2013. *Technology NOW*. https://philwordexpress.ph/

[20] Saroj, Y. and Chaudhary, T. 2015 *Study on Various Clustering Techniques*. www.ijcsit.com/

[21] Sembiring, R.W., Zain, J.M. and Embong, A. 2010. *A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course*. http://www.studymode.com/

[22] Sigwai, R. 2010. *Databases*. www.springer.com/

[23] Sikes, N. 2015. *Resolving Big Data*. http://atinta.wmwikis.net/

[24] Sondwale, P. 2015. *Overview of Predictive and Descriptive Data Mining Techniques*. https://www.ijarcsse.com/

[25] Takagi, D. 2013. *Data Evaluation, Vol. 1: Processing Systems*. University of Technological Solutions. North-Holland.

[26] Terro, L.C. 2016. *What Is Clustering?* www.libguides.williams.edu/

[27] Wihls, J. 2014. *Communication*. https://www.academia.edu/

[28] Wilsen, J. 2014. *Powerful Tool: Data Mining Clustering Technique*. http://www.papercamp.com/

[29] Zaiane, O.R. 2013. *Data Mining*. https://webdocs.cs.ualberta.ca/