Proceedings of the Second

University Journal of Research and Innovation

August, 2020

Organized by

University of Computer Studies (Pakokku)

Proceedings of

The Second University Journal of Research and Innovation 2020

Augest, 2020

Organized by

University of Computer Studies (Pakokku)

Department of Higher Education,

Ministry of Education, Myanmar

University Journal of Research and Innovation

Volume 2, Issue 1 2020

University Journal of Research and Innovation 2020

Volume 2, Issue 1, 2020

This journal and individual paper published at www.ucspkku.edu.mm.

Copies:110

All research papers in this journal have undergone rigorous peer-reviewed which is published annually. Full papers submitted for publication are refereed by the Associate Editorial Board through an anonymous referee process.

The authors of the paper bear the responsibility for their content.

Papers presented at the Second University Journal of Research and Innovation (UJRI), University of Computer Studies (Pakokku), August 2020.

UJRI 2020 Editorial Board

UJRI 2020 Editorial Board

Proceedings of

The Second University Journal of

Information and Computing Science 2020

Augest, 2020

Contents

Database Management System & Information Retrieval

Digital Signal Processing

Embedded System

Image Processing

Network & Security

Software Engineering and Web Engineering

Assessment of Teachers' Performance Factors by Using K-Means Clustering

May Su Hlaing
University of Computer Studies (Pakokku)
maysulay84@gmail.com

Shwe Sin Thein
University of Computer Studies (Pakokku)
shwesinthein@ucspkku.edu.mm

Abstract

In an academic Institution or University, qualified teachers are needed to promote the students' academic outcomes and improve the standard of education. The principal or administrator should evaluate and monitor the teachers' performance in higher Institutions. Therefore, the principal investigates the assessment factors related to the evaluation of teachers teaching performance using predictive data mining techniques. We got these factors from the questionnaire or interview with a related person such as principals, teachers, and other educators. This proposed paper presents the relevant solutions for educational and administrative problems to improve the evaluation of teachers' performance by applying data mining techniques in higher Institutions. This study is intended to propose and predict the teachers' performance by using the K-means clustering algorithm in data mining. We collect the teacher's data set from the University of the Computer Studies (Pakokku and Mandalay).

Keywords: Academic Outcomes, Assessment, Higher Institution, Evaluations, K-Means

1. Introduction

In the current age, data mining is a very interesting field in the data analysis area. It is gradually recognized as a new tool to extract a valuable and meaningful pattern from a large data set. The extracted pattern may be used to predict future manner. The researchers applied these data mining techniques to solve the problems in different domains such as health care, industry, business, financial, and education in academics.

This paper deeply describes the assessment of educators' or teacher's performance to apply data mining techniques in the educational domain. In order to classify and predict teachers' performance in educational institutes, we survey and collect the facts that highlight teachers' performance.

There are different faculties in University or Institute. The nature of performance factors is so many different things between these faculties. For example, Myanmar's subject has no practical method for Lab. But Computer Graphic subject has to be done in the computer laboratory room with computer hardware and software. So, the performance factors of teachers from these two subjects cannot be evaluated based on the same questionnaires. Thus, we find the common factors for a different level of teachers in various faculties.

2. Related Works

There are many kinds of researches in this area. That is leading toward improving the performance of teachers and improve their courses in the educational process.

The author Ajay Kumar Pal et al [1] used data mining algorithms such as Naive Bayes, ID3, CART, LAD to prove the best algorithm based on the data.

Ahmed Mohamed [2] used J48 Decision Tree, Multilayer Perception, Naïve Bayes, and Sequential Minimal Optimization algorithms. They observed that a comparison of all the four classifiers is conducted to predict the accuracy and to find the best performing classification algorithm among all.

K. Devasenapathy and S Duraisamy [3] used Iterative Dichotomies 3(ID3) algorithms in data mining methods. They surveyed that ID3 supports the Institute to grow the performance.

The author Oswal Sangita and Jagli Dhanamma [4] used K-Means clustering algorithm. They described the data acquirement, Cluster formation and analysis on how to obtain a useful teaching evaluation data and created an appropriate cluster.

The author Yi Hua [5] used Data mining Algorithms (Range, Quartiles, Variance, Standard Deviation,) to study that school administrators should focus on the structure and associated evaluation indicators of performance pay and findings suggest that local governments should increase funding in teacher performance pay if it is to be successful.

The author Simon Borg, Ian Clifford, Khing Phyu Htut [6] used datamining method Teaching Evaluation indicator. Their analysis has highlighted factors which facilitated the positive impact of the teaching and teacher education.

We mainly focused on predicting the performance of teachers in academics at our University using the clustering algorithm such as K-means clustering algorithm in this paper.

3. Data Mining Techniques in Education

Data mining methodologies are used in many application domains such as marketing, medicine, finance, management, and recently in education that is known as educational data mining.

A lot of observations in research emphasized enhancing the performance of teachers and improves the courses and curriculum.

Educational data mining (EDM) deals with developing methods for exploring data from educational sectors with the purpose of providing quality education to students and to make effective managerial decisions [7].

By offering precisely directed courses to the teacher according to his need and build on what he has from previous knowledge [8].

The purpose of this paper is to assess teachers' performance through the study of their specialization and expertise and the year of service of the educational process, evaluate and determine courses for teachers under improving their performance. This study will provide great support to the head of the department for decision-making in educational institutions. It will provide a framework for assessing teachers' performance relating to the adopted performance criteria.

This research is to develop a model based on data mining that evaluates the performance of teachers, apply our approach on real data sets. This paper investigates the educational domain of data mining using a case study from the teacher data collected from the University of Computer Studies (Pakokku) and the University of Computer Studies (Mandalay). We collected the teachers' data set in the period [2019-2020]. It showed how could we preprocess the data, how to apply data mining methods on the data, and finally how can we benefit from the discovered knowledge. There are many kinds of knowledge that can be discovered from the data. In this work, we investigated the most common ones which are association rules, classification. WEKA tool is used for applying the methods on the teacher's data set.

4. Clustering

Clustering is a method to group data into classes with identical characteristics in which the similarity of intra-class is maximized or minimized. Clustering is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes.

4.1. K-means Clustering Algorithms

K-means is one of the simplest unsupervised learning algorithms used for clustering. K-means partitions n observations into k clusters in which each observation belongs to the cluster with the nearest mean. K-means was used to cluster individual students based on their online activities to reveal information that was missing from team-wise clustering [4].

This algorithm aims at minimizing an objective function, in this case, a squared error function. The algorithm aims to minimize the objective function. K-means is one of the simplest unsupervised learning algorithms used for clustering. K-means partitions n observations into k clusters in which each observation belongs to the cluster with the nearest mean (equation (1) & (2)). This algorithm aims at minimizing an objective function, in this case, a squared error function.

$$E = \sum_{i=1}^{k} \sum_{p \in c_1} dist(p, c_i)^2 \quad \text{-------(1)}$$

Euclidean Distance $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ ------- (2)

where E is the sum of the squared error for all objects in the data set; $\boldsymbol{p}$ is the point in space representing a given object, and $\boldsymbol{c_i}$ is the centroid of cluster C_i (both $\boldsymbol{p}$ and $\boldsymbol{c_i}$ are multidimensional).

5. Constructing Teacher Dataset

In this section, we present three portions to construct the teacher data set. Firstly, we collect the teacher raw data by interviewing experts, learning by being told, and learning by observation. And then, we identify feature and label sources by interviewing experts. Finally, we select a sampling strategy (stratified sample) according to the designation.

The data set used in this paper contains the teacher's information collected from the University of Computer Studies (Pakokku) and the University of Computer Studies (Mandalay) in the period 2019-2020. The Teacher data set consists of 130 records and 21 attributes after combining the training, administrative, and questionnaire information for the best training. Table 1 presents the attributes and their description that exists in the data set as taken from the data source. In Table 1, we choose attributes (mark as √) from this table. The selected attributes can be evaluated as factors on the performance of teachers.

Table 1. Attributes & description for data set

Sr.no	Attribute	Description	Selected
1	name	Teacher's name	
2	Position	Teacher' rank {T, AL, L, AP, P}	✓
3	Professionalism	Work skilful	✓
4	Customer Focus	Responsiveness of the teachers to; students/pupils, parents	✓
5	Integrity	the teacher exhibits honesty,	✓
6	Team Spirit	consider the ability of the teacher to work in a team	✓
7	Innovativeness	to introduce new ideas	✓
8	Department	{SW, IS, ITSM, FOC, NS, LD, FOCST}	
9	Academic Qualification Degree	Qualified degree	✓
10	Age	The age of teacher	
11	Services	Working years	✓
12	Awards	Professional ranking	✓
13	Seminar	Celebrating conference, micro teaching	✓
14	Training	Training different subjects or courses	✓
15	Lab	Practical time	✓
16	Publications	No of papers have been published	✓
17	Teaching in specific subjects	no. of subject, subject name	✓
18	Serving in other duties	no. of activities	✓
19	Working with students outside of class time	Teacher's morality for working load with students	✓
20	Teaching material	Teaching method	✓
21	Class	Teacher's quality {Excellent, Moderate, Good}	✓

5.1. Data Acquisition Form

In this section, we present the data acquisition form from attaining by interviewing or questionaries' facts with principals and teachers.

In the following figures (Figure 1 and Figure 2), it represents that form 1 refers collection of data filling themselves (teachers) and, form 2 refers to gathering data from the principal and faculty dean on each teacher. We specify the five categories for form 2: 1 -bad, 2 - moderate, 3-good, 4-very good, 5 - excellent.

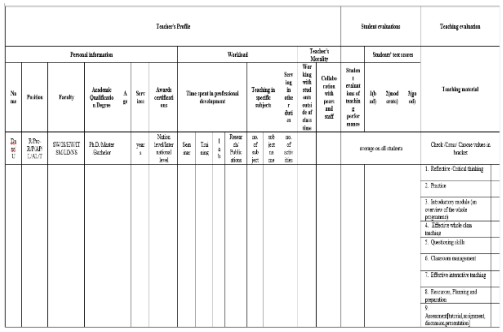

Figure 1. Data acquisition form for each teacher (form 1)

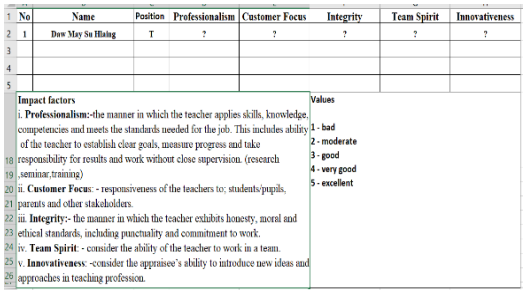

Figure 2. Data acquisition form for each teacher by principal, faculty dean (form 2)

6. Experimental Results and Observation

After we got the raw data with form, we transform the data format with CSV (comma-separated values) form. A comma-separated values file is a text file that uses a comma to separate values. Each line of the file is a data record. We use Excel to change from tabular data to CSV format. Table 2 shows the sample data set. Then, we use the WEKA tool from CSV format to Arff format in Figure 3.

The dataset consisted of 130 records. The data for all assessments were transformed into proper normal forms appropriate for mining. Normalization was done on these attributes so that data should reduce in a small specified range not to outweigh the measurement of other attributes.

Table 2. Teacher's dataset (sample records 130)

sr.no	Position	Professio -nalism	Customer Focus	Integrity	Team Spirit	Innovati -veness	Faculty	Academic Qualification Degree	Age	Services
1	T	4	4	3	4	3	SW	Master	29	2
2	T	3	3	2	2	3	IS	Master	30	2
3	T	3	2	2	2	2	IS	Master	28	2
4	T	3	3	3	4	3	ITSM	Master	33	2
5	T	3	4	3	3	4	ITSM	Master	31	2
6	T	3	3	3	3	2	FOC	Master	29	2
7	T	2	3	3	3	3	NS	Master	27	2
8	T	3	3	3	3	2	LD	Master	29	2
9	T	3	3	3	3	2	LD	Master	27	2
10	AL	4	4	3	4	2	SW	Master	30	5
11	AL	3	3	2	2	3	IS	Master	28	5
12	AL	3	3	3	3	2	FOC	Master	31	6
93	T	3	3	3	3	2	LD	Master	25	2
94	L	3	3	4	4	3	FOCST	Master	34	11
125	AL	4	4	3	4	4	ITSM	Master	41	16
126	T	3	3	3	3	3	ITSM	Bachelor	35	11
127	P	4	5	4	4	4	ITSM	PhD	41	18
128	AP	4	5	3	4	4	FOC	Master	48	19
129	AL	4	4	3	4	4	FOC	Master	35	10
130	P	4	5	3	4	4	FOC	PhD	47	21

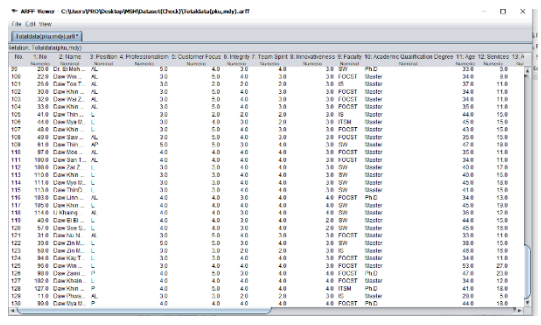

Figure 3. Open arff file format in WEKA

6.1. Preprocessing Data Set

Teacher academic performance is predicted based on several input attributes. An Algorithm such as K-means clustering was used on the input attributes to generate a model to the predict academic performance of teachers. In this research, WEKA software was used for preprocessing input data.

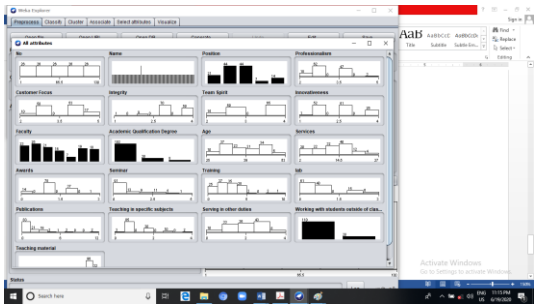

Figure 4. Data preprocessing in WEKA

Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data (Figure 4).

6.2. Results of K-means Algorithm

After data preprocessing, we have to cluster the data set for the class label. So, this section presents the results of the K-means clustering algorithm with WEKA as follows.

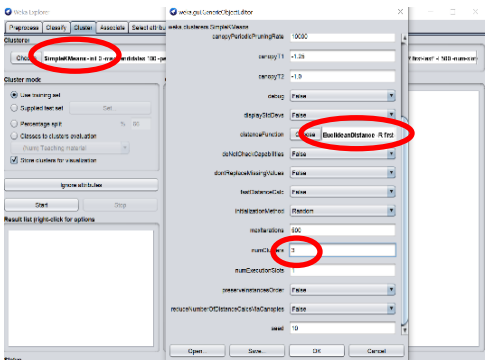

Figure 5. Clustering in WEKA

We choose the cluster tab in WEKA and then choose simple K-means cluster and define cluster number as we want to cluster.

In Figure 5, we put 3 into the "numCluster" text box. So it evaluates and clustering the data set as three clusters for labeling the class.

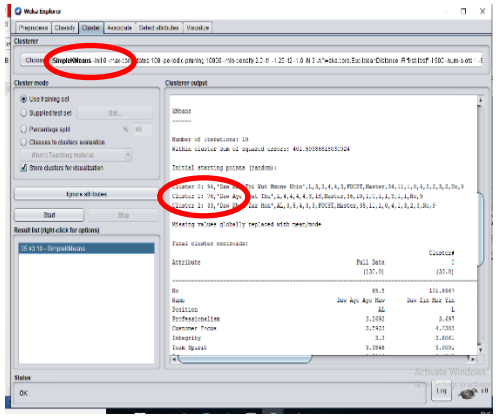

Figure 6. Clustering output

After using the K-means clustering algorithm, the clustering output is resulted as in Figure 6.

The results from Table 3 show that there are three clusters in K-means. In K-means clustering the third cluster has maximum accuracy. The first and second clusters have less accuracy. The time taken by the K-means algorithm to build model is 0.02 second and number of iterations is 10.

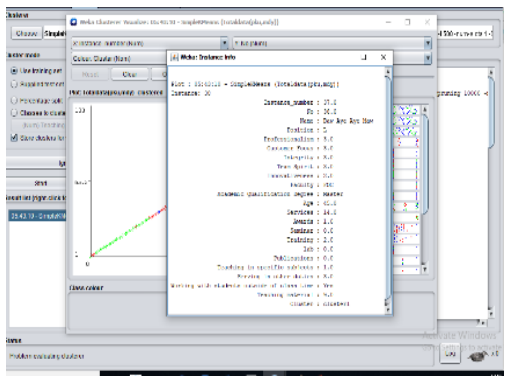

Figure 7. Visualizing for output

Table 3. Clustering results for k-means

Cluster #	k-means cluster size	k-means percentage
First (cluster 0)	33	25%
Second (cluster 1)	42	32%
Third (cluster 2)	55	42%

Finally, we cluster the teacher data set in academic institutes. We have to label the class on each data as follows.

Initial starting points (random):
Cluster 0: **AL**,4,4,4,4,3, **Master**,**11**,1,0,3,1,3,1,1, No,9 = "**Moderate**"
Cluster 1: **P**,4,4,4,4,4, **Ph. D**,**19**,2,0,1,0,5,1,1, Yes, 9 = "**Excellent**"
Cluster 2: **L**,4,4,3,4,2, **Master**,**16**,1,0,2,1,2,2,2, No,10 = "**Good**"

We define the class according to the point of view of the experts, look at the centroid value of data from the output of the WEKA and nature of data. Based on these factors there are three hypotheses (Excellent, Good, Moderate) that will be defined.

So, we guess these data sets to assess the teachers' performance based on their position, degree, and experience (years of service). As describe in clusters, the highest level of position, highest degree, and more experience in service for a teacher will be got the great class "Excellent". In this way, we can label the class for all data set.

In addition, this research supports all administrators or faculty dean to make the best decision for their teachers. They can know how they will continue to instruct or train their teachers based on their performance.

6.3. Benefits of Proposed System

We want to explain how to impact the teachers' improvement according to the performance factors as follows.

This system is necessary for the institutions to manage and develop the skills of teachers. This helps the faculty dean to make the important decision by his/her teachers' performance. If the teachers' quality or output may not good, the dean has to arrange to train them based on their subjects.

Research efforts are also needed in the overall development of faculty members to improve the quality of teaching and learning process.

The outcome can be used to find the steps needed for further improvement of their performance. The proposed system will assist the higher administrators in making appropriate managerial decisions, which will optimize the academic outcomes and improves the standard of education.

7. Conclusion

In this research work, clustering algorithms were examined based on the teachers' data set via its attribute values. The clustering algorithm provides well in the prediction of teachers' performance. This research can be sustained further by adding different teachers' attributes that have an impact on academic performance, which are not used in this data set. Also, the size of the data set may be increased such that the teacher data from other Universities is used instead of data from a few Institutes in a particular region to grow the accuracy of results. In this paper, we have done the data set from two Computer Universities (UCS (Pakokku) and UCS (Mandalay)).

We observed that **Position**, **Degree**, **Service** are the most important factors to qualify each teacher's performance.

References

[1] Ajay Kumar Pal, Saurabh Pal, "Evaluation of Teacher's Performance: A Data Mining Approach", International Journal of Computer Science and Mobile Computing 2018.

[2] Ahmed Mohamed, Ahmet Rizaner, Ali Hakan UIusoy, "Using data Mining to Predict Instructor Performance" Article in Procedia Computer Science, 2016.

[3] K. Devasenapathy, S. Duraisamy, "Evaluating the Performance of Teaching Assistant Using Decision Tree ID3

Algorithm," International Journal of Computer Applications, 2017.
[4] Oswal Sangita and Jagli Dhanamma "An Improved K-Means Clustering Approach for Teaching Evaluation" Vivekanand Education Society's Institute of Technology, 2011.
[5] Yi Hua, "Teacher Perceptions of Teacher Performance Pay and Performance Evaluation in Yunnan Province, China" (2017).
[6] SIMON BROG, Ian Clifford, Khaing Phyu Htut "Teaching and Teacher Education" Western Norway University of Applied Sciences, 2018
[7] Ms. Aanchal K Patil1, Prof. S. R. Nagarmunoli "Instructor's Performance Evaluation System using Data Mining Techniques" International Journal of Advanced Research in Computer and Communication Engineering, 2017.
[8] Randa Kh. Hemaid1, Alaa M. El-Halees2 "Improving Teacher Performance using Data Mining" International Journal of Advanced Research in Computer and Communication Engineering, 2015.

The Use of ICTs in Teaching-Learning-Assessments: A Study in University of Computer Studies (Pakokku)

San San Nwel, Su Mon Han, Thet Thet Aye Mon
University of Computer Studies (Pakokku)
sansannwel@ucspkku.edu.mm

Abstract

Information Communication Technologies are the power that has changed many aspects of the lives. Many countries now understand the importance of ICT and mastering the basic skills and concepts of it as part of the core of education. Assessment is a key component of teaching and learning because it helps both teachers and students. For teachers, assessment information is used to adjust their teaching strategies. The students can adjust their learning strategies by using assessment information. The main aim of this paper is to prepare teaching and learning assessments to meet the requirements of 21 century for University. We analyze the different types of assessment methods for each subject by using Grounded Theory. In this paper, the empirical results of analysis about the use of ICTs for assessment in University of Computer Studies (Pakokku) in 2018-2019 academic year are presented. Those empirical results support the extent of ICT usage, assessment types that should be used for evaluation and control of budget plan to support ICT resources for each faculty and department.

Keywords: education, ICTs, assessment process, teaching strategies, learning strategies.

1. Introduction

Information and communication technologies (ICTs) have become commonplace entities in all aspects of life. The use of ICTs in education lends itself to more student-cantered learning settings. But it often creates some tensions for some teachers and students [1]. With the world moving rapidly into digital media and information, the role of ICTs in teaching and learning is becoming more and more important. This importance will continue to grow and develop in the 21st century [2].

Today, universities and colleges need to assess their subjects to improve the quality of teachers and students. Assessment allows teachers to see if their teaching has been effective. Assessment also allows teachers to ensure students learn what they need to know in order to meet the course's learning objectives. For students, they are able to see how they are doing in a class; they are able to determine whether or not they understand course material. Thus, assessment can also help to motivate students [4].

In this paper, the assessment types such as project, lab test, practical, Moodle test, language lab and 4 skills are specified that the subjects used ICTs. But the subjects used the assessment types such as assignment and tutorial are specified not use of ICTs. Then the use of ICTs is calculated by dividing the number of subjects used ICTs by the total number of subjects.

2. Information and Communication Technologies (ICTs)

ICTs are an acronym that stands for "Information and Communication Technologies". Information and communication technologies are an umbrella term that includes all technologies for the manipulation and communication of information. ICTs consider all the uses of digital technology that already exists to help individuals, business and organization. It

is difficult to define ICTs because it is difficult to keep up the changes, they happen so fast. ICTs are concern with the storage, retrieval, manipulation, transmission or receipt of digital data. ICTs are also defined as "ICTs is the computing and communication facilities and features that variously support teaching, learning and a range of activities in education" [3].

Teaching and learning process is always going together; we cannot consider these two as separate and independent activities. In fact, these are similar as two sides of the same coin, interconnected and interrelated. The process of teaching and learning in institutes around the world can be divided into four main stages. These four stages are shown in Table 1 [5].

Table. 1 Model of stages of teaching and learning using ICT

Stage 1	Discovering ICT tools
Stage 2	Learning how to use ICT tools
Stage 3	Understanding how and when to use ICT tools to achieve particular purposes
Stage 4	Specialization in the use of ICT tools

As shown in Table.1, firstly teachers discover the appropriate ICT tools for their teaching-learning-assessment according to their subject' nature. Then they learn how to use ICT tools. And teachers try to understand how and when to use ICT tools to achieve particular purpose. Finally, they specialize to use ICT tools in their teaching-learning-assessment.

3. Grounded Theory

Grounded theory involves the collection and analysis of data. The theory is "grounded" in actual data and Grounded theory provides qualitative researchers with guidelines for collecting and analyzing data. Grounded theory commonly uses the following data collection methods [7]:

- Interviewing participants with open-ended questions.
- Participant Observation (fieldwork) and/or focus groups.
- Study of Artifacts and Texts

Grounded theory can be used with different types of data, including: numerical data, interview transcripts, focus group transcripts, literature, and, more recently, contemporary data sources such as video/DVD recordings, websites, and secondary datasets. Grounded Theory is indeed widely used in areas that are considered more exploratory or discovery-oriented. It is a general method that can be applied to very little research [6]. Grounded Theory is not only simply a method for analyzing interview or focus group data but also that it informs all aspects of the design and implementation of the study [8].

In this proposed system, Participant Observation (fieldwork) and/or focus groups method and Study of Artifacts and Texts method are used to collect data such as number of subjects, subject code, subject names, type of assessment methods for each faculty and department.

4. Assessment Process Cycle

In assessment process cycle, the head of faculty, head of department and classroom teacher define what they want student to learn firstly. Then, they design appropriate learning outcomes using level descriptors. And then, they decide how student can best show they have achieved these learning outcomes. Then, the most appropriate method is chosen. They design assessment criteria and feedback format based on learning outcomes and finally engage in a dialogue with students about feedback and what they want student to learn.

As shown in Figure 2, in our University, teachers need to check their students with assessment tasks whether teaching and learning activities meet their learning outcome. Teachers need to consider and plan the types of assessment methods and marking criteria and then review feedback of their students. According to the feedback, if needed our University decides to

change current curriculum and teaching strategy. In order to support these teaching and learning activities, types of assessment method and feedback, data collection methods are used. By using Moodle for assessments task, both teachers and students get many benefits such as time consuming, getting exact results, preventing error prone, storing result scores in Moodle database and progressing ICT skills.

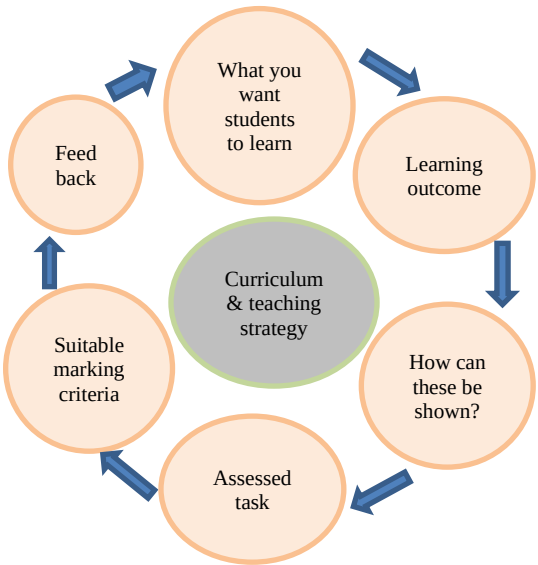

Figure 1. Assessment Process Cycle

5. Empirical Results

In our University, there are four faculties and three departments in teaching. They are Faculty of Computer Science, Faculty of Computer Science and Technology, Faculty of Information Science, Faculty of Computing, Information Technology Supports and Maintenance department, Natural Science department and Language department. Number of subjects lectured by each faculty and department and the use of different assessment methods for each subject are shown in the following tables.

Nowadays, for Higher Education the role of ICT is very important. Thus, our University needs to investigate to what extent ICT is used for teaching-learning-assessment. We need to study what subjects should use ICT for their assessments. Our empirical results explore the extent of ICT usage for each faculty and department.

5.1. Faculty of Computer Science (FCS)

In University of Computer Studies (Pakokku), Faculty of Computer Science plays the important role in teaching. There are fourteen subjects lectured by this faculty. Table.2 and Figure 2 shows the different assessment method applied to assess the subjects for different courses and the use of ICTs for teaching-learning-assessment respectively.

Table. 2 Subjects and their assessment methods

No	Subject	Types of Assessment Method					
		Project	Lab test	Practical	Moodle test	Tutorial	Assignment
1	CST-1112 Principle of IT	-	√	√	√	√	√
2	CST-1212 Programming Logic & Design	-	-	√	-	√	√
3	CST-2112 Principle Computer Science II	-	-	√	√	√	√
4	CST-2212 Data Structure & Algorithm	-	-	√	√	√	√
5	CST-2215 J2EE Programming	√	-	√	√	-	√
6	CS-3212 Operating System	-	-	√	√	√	√

7	CS-3106 Advanced Programming Technique	-	-	-	√	√	√
8	CS-3216 Professional Ethics in IT	√	-	-	√	√	√
9	CST-4101 Artificial Intelligence	√	-	-	-	√	√
10	CST-4103 Analysis of Algorithm	√	-	-	-	√	√
11	CS-4213 Operating System	-	-	√	√	√	√
12	CS-4216 Computer Graphic	-	-	√	√	√	√
13	CS-5102 Distributed System	√	-	-	-	√	√
14	CS-5104 Computer Graphic	√	-	-	-	√	√

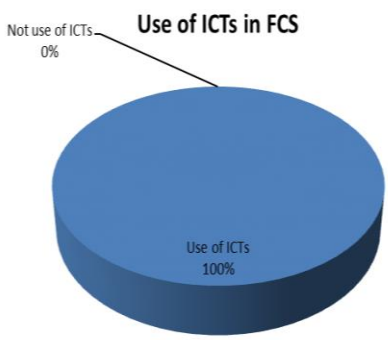

Figure 2. Use of ICTs in FCS

In this faculty, there are fourteen subjects lectured to each course. All these subjects use at least one types of assessment method such as project, lab test, practical or Moodle test. The use of ICTs for this faculty is calculated by dividing the number of subjects used ICTs (project, lab test, practical or Moodle test) by total number of subjects.

5.2. Faculty of Information Science (FIS)

Faculty of Information Science performs a vital role of our university. This faculty lectures twelve subjects for different courses. For these subjects, many different assessment methods are applied for teaching-learning-assessment. For this faculty, the use of ICTs to assess subjects is shown in Figure 3.

Table 3. Subjects and their assessment methods

No	Subject	Types of Assessment Method					
		Project	Lab test	Practical	Moodle test	Tutorial	Assignment
1	CST-2124 DBMS	√	-	√	√	-	√
2	CS-3124 SE	-	-	√	√	√	-
3	CS-4125 SE	√	-	-	√	√	-
4	CS-4125 DBMS	√	-	√	√	-	√
5	CS-5103 IAS	√	-	√	√	√	-
6	CS-5105 Data Mining	√	-	√	√	-	√
7	IS-101 SE	√	-	√	√	-	√
8	CS-304 DBMS	√	√	-	-	√	√
9	CT-401 DBMS	√	√	-	√		√
10	CS-404 MIS	-	-	-	√	√	-
11	CS-404 IAS	-	-	-	√	√	√
12	CS-405 UML	√	√	-	√	-	√

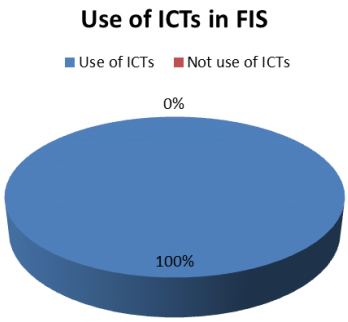

Figure 3. Use of ICTs in FIS

5.3. Faculty of Computer Science and Technology (FCST)

In teaching, Faculty of Computer Science and Technology is a main source for computer hardware. There are twenty-five subjects to lecture to students. For these subjects, the different assessment methods are used to assess them. All subjects apply ICTs for their assessments and the use of ICTs for this faculty is shown inf Figure 4.

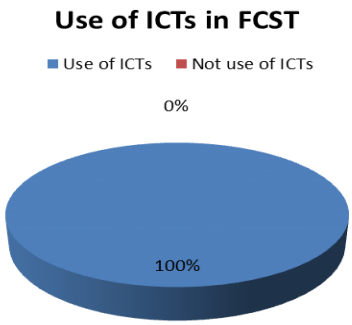

Figure 4. Use of ICTs in FCST

5.4. Faculty of Computing (FC)

In our university, Faculty of Computing supports students with computational thinking, logic and concept. This faculty delivers seven subjects to students in different courses. All subjects use ICTs to assess tasks and the use of ICTs for assessment for different course is as shown in Figure 5.

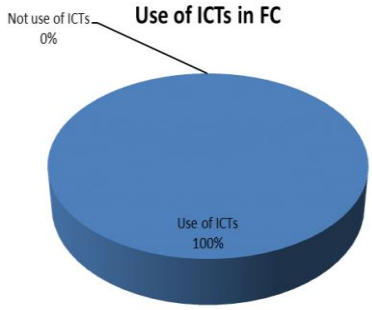

Figure 5. Use of ICTs in FC

5.5. Information Technology Supports and Maintenance (ITSM) Department

In our university, ITSM Department supports students to build systems and applications that help to solve today's problems and share skills and knowledge with the community. There are six subjects lectured by this department and among them, five subjects use ICTs for assessments. Figure 6 shows the use of ICTs to assess these subjects.

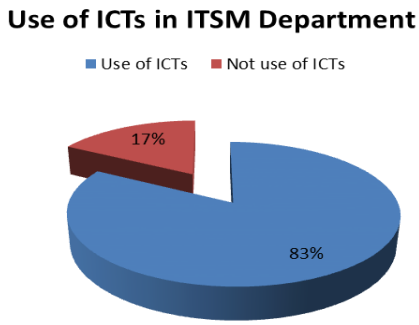

Figure 6. Use of ICTs in ITSM Department

5.6. Natural Science Department (NS)

Natural Science Department supports the students with the methods scientists used to explore natural phenomena, including observation, hypothesis development, measurement and data collection,

experimentation, evaluation of evidence, and employment of mathematical analysis. In this department, practical are demonstrated by using laboratorial instruments but ICTs are not used for this subject. Therefore, the use of ICTs for this department is shown in Figure 7.

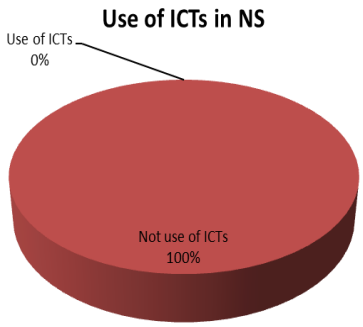

Figure 7. Use of ICTs in NS Department

5.7. Language Department

In teaching, Language Department provides both English language and Myanmar language. For students, it is essential not only 4 skills to chance opportunities for living but also civilization and patriotism to develop our country. There are six subjects lectured by this department and among them, five subjects use ICTs for their assessments. However, the subject for Myanmar is not used ICTs to assess tasks. Figure 8 reveals the use of ICTs for these department.

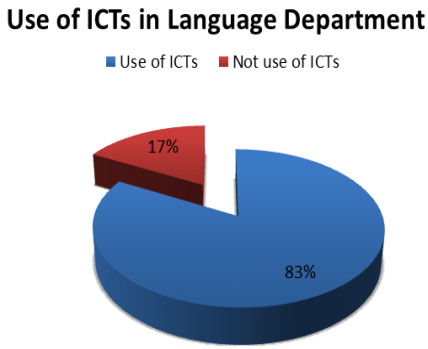

Figure 8. Use of ICTs in Language Department

In this paper, we consider the assessment types such as language lab and 4 skills use ICTs because the student and teacher access to information, promote interaction and communication with Bluetooth speaker for audio files.

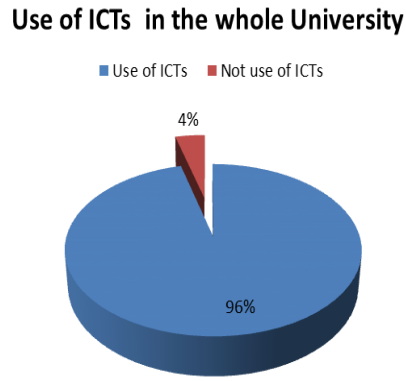

Figure 9. Use of ICTs in the whole University

In our University, there are 71 subjects lectured by all faculties and departments for different courses. Among these subjects, only three subjects such as marketing concept, physics and Myanmar are not assessed for assessment by using ICTs because of their nature. On the whole, the use of ICTs in our University is 96% as shown in Figure 9.

6. Conclusion

The present age is the age of technology, whereby technology plays a key role in daily lives; this also includes the education system. There are endless possibilities with the integration of ICT in the education system. The use of ICT in education not only improves classroom teaching - learning - assessment processes, but also provides the facility of e-learning. The total numbers of subjects lectured by our University is 71 subjects. Among them, 68 subjects use ICTs for teaching-learning-assessment. For teaching-learning-assessment process, the use of ICTs in the whole University is 96% in 2018-2019 academic years.

To investigate data in real-life context, using Grounded Theory can identify contingent nature of practice and is better at determining what actually happen. This paper reveals the use of ICT in real-life situation of each faculty and department of our University. Consequently, if we know the ICT usage of faculties and departments, our University can manage budget share used to support ICT resources effectively and efficiently.

References

[1] https://learntechit.com

[2] https://www.researchgate.net

[3] ICT in Education (2006), Information and communication technologies in teacher education: A planning guide, 2007.

[4] Kameron, Saskia E., "A Review of Free Online Learning Management Systems (LMS)", TESL-EJ, ISSN 1872-4303, vol.7, No.2, M-2, http://www-writing.berkeley.edu/TESL EJ/ej26/m2.html, 2003.

[5] Branzburg, Jeffrey "How To: Use the Moodle Course Management System", http://www.techlearning.com/story/sho, Aug 15, 2005.

[6] Goede, R., and Villiers C, De. "The Applicability of Grounded Theory as Research Methodology in studies on the use of Methodologies in IS Practices", Proceedings of SAICSIT, South Africa, 2003, pp. 208-217.

[7] Allan, G. "A critique of using grounded theory as a research method," Electronic Journal of Business Research Methods (2:1), 2003, pp. 1-10.

[8] Hansen, Bo H., and Kautz, K. "Grounded Theory Applied –Studying Information Systems development Methodologies in Practice", Proceedings of the 38th Hawaii International Conference on System Sciences, Hawaii, US, 2005.

Movies Borrowing Analysis using Closet Algorithm

Seint Wint Thu
University of Computer Studies, Meiktila
seintwint241@gmail.com

Pa Pa Win
University of Computer Studies, Meiktila
papawin.pale@gmail.com

Zin Mar Naing
University of Computer Studies, Meiktila
zinmarnaing2016mtla@gmail.com

Abstract

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Association rule mining is an important role in data mining to extract the associated itemsets in transactional database and examine user behaviors. The CLOSET algorithm was designed to extract frequent closed itemsets from large databases. CLOSET is an FP-tree-based database projection method for efficient mining of frequent closed itemset. The system generates the frequent closed movie-sets and association rules. The system generates strong association rules when the minimum support count and confidence square measure high. This paper provides that borrowers are interested which movies, actors and actresses according to the generated strong rules.

Keywords: Data mining, CLOSET, association rules, FP-tree-based database projection

1. Introduction

Data Mining is the process of discovering new correlations, patterns, and trends by digging into large amounts of data stored in warehouses. Data mining is an analytical tool for analyzing data. It allows users to analyze data, categorize it, and summarize the relationships among data. Technically, data mining is the process of finding correlations of patterns in large relational databases. It involves some common tasks like clustering, association rule mining, regression, and classification etc. Among them, association rule mining searches for relationships between variables. Mining complete set of itemsets suffers from generating a very large number of itemsets and association rules. With the help of the association rule, the borrowers can know which movies are frequently interested. This information generated from the strong association rule provides users to borrow the interested movies. Data mining is used to present the mined knowledge to the user behavior [1].

Data mining has attracted a good deal of attention within the information business and in society. The information and knowledge gained can be used for applications starting from market research, fraud detection, and customer retention. The rapid growth and integration of databases provides scientists, engineers, and business individuals with a vast new resource to analyze scientific discoveries. It is used to optimize industrial systems, and uncover financially valuable patterns [2].

Out of this research has come a wide variety of learning techniques. They have the potential to transform many scientific and industrial fields. Several research communities have converged on a common set of issues surrounding supervised, unsupervised, and reinforcement learning issues [7].

2. Related Works

Data mining is also known as extracting or mining information from data set. The frequent itemsets is the important part of data mining. So, some concepts are required to find out frequent itemsets. J. Pei et al. proposed an efficient method for mining frequent closed itemsets without candidate generation in [3], called CLOSET.

Frequent pattern mining in data classification has become an important data mining task [4]. These patterns can be represented in the form of association rules. Support and confidence of association analysis are two measures of interesting rules. Association rules are interesting if it is satisfy both a minimum support threshold and minimum confidence threshold. An alternative method is to identify the closed itemsets. This method uses the minimum support to determine which ones are frequent.

This paper presented that CLOSET algorithm discovers interesting association or correlation relationships among large number of data items based on correlation analysis.

3. Association Rule Mining

Association rule mining approach discovers interesting relations between variables in large database. It also discovers frequent itemsets. Association rule mining is one of the data mining techniques. It is designed to group objects together from large database aiming to extract the interesting correlation and relation among large amount of data. It is intended to identify strong rules discovered in databases using some measures of interestingness. It is widely used to discover correlations in transactional data. Association rules are if/then statements that help to uncover relationships between unrelated data in a database. Association rules are used to find the relationships between the objects which are frequently used together. For example, if the customer buys laptop then he may also buy memory card.

There are two basic criteria that association rules use, support and confidence. It identifies the relationships and rules generated by analyzing data. The support is simply the number of transactions in which the itemsets occur. The support is sometimes expressed as a percentage of the total number of records in the database. Confidence is the conditional probability of occurrence of consequent given the antecedent. Lift can be used to compare confidence with expected confidence. Lift is one more parameter of interest in the association analysis. It is a frequent itemset that is both closed and its support is greater than or equal to minimum support. Association rules are usually needed to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time [5].

Rule: A => B --------(1)

Support = frq (A,B) / N --------(2)

Confidence = frq (A,B) / frq (A) -------(3)

4. CLOSET Algorithm

CLOSET algorithm is a pattern growth method providing on dataset organization. CLOSET is economical and scalable over large database. CLOSET algorithm uses the principles of the FP-Tree data structure. It is used to avoid the candidate generation step during the process of mining frequent closed itemsets. CLOSET has three developed techniques for mining closed itemsets : (1) applying a compressed, frequent pattern tree FP-tree structure for mining closed itemsets without candidate generation, (2) developing a single prefix path compression technique to identify frequent closed itemsets quickly, and (3) exploring a partition-based projection mechanism for scalable mining in large database.

The CLOSET approach is mainly identified for its efficiency. The optimization method of using CLOSET algorithm is to extract every item appearing in the conditional databases of the frequent item subsets. This reduces the size of the FP-tree. This improves the overall speed of the recursive process by combining some items. There are several advantages of CLOSET over other approaches:

- It constructs a highly compact FP-tree based on conditional databases.
- It avoids costly candidate generation and test by successively concatenating frequent 1-itemset found in the (conditional) FP-trees.
- It applies a partitioning-based divide-and-conquer method which dramatically reduces the size of the subsequent conditional pattern bases and conditional FP-tree.
- It develops a single prefix path compression technique to identify frequent closed itemsets quickly.

The CLOSET algorithm is very efficient. But it is a highly complex technique. The mining procedure of CLOSET follows the FP-growth algorithm. CLOSET treats items appearing in every transaction of the conditional database specially. The algorithm extracts only the closed patterns by careful movie-keeping. For example, if Q is the set of items, it appears in every transaction of the P conditional database. Then, P ∪ Q creates a frequent closed itemset if it is not a proper subset of any frequent closed itemset with the equal support. CLOSET also prunes the search space. For example, if P and Q are frequent itemset with the equal support, where Q is also a closed itemset. P ⊂ Q, then it does not mine the conditional database of P because the latter will not produce any frequent closed itemsets [6].

Algorithm (CLOSET): Mining frequent closed itemsets by the FP-tree method
Input: Transaction dataset TDB and support threshold old min_sup;
Output: The complete set of frequent closed itemsets;
Method: 1. Initialization. Let FCI be the set of frequent closed itemset. Initialize FCI□∅; 2. Find frequent items. Scan transaction database TDB, compute frequent item list f-list; 3. Mine frequent closed itemsets recursively. Call CLOSET (∅, TDB, f-list, FCI).

4.1. Sample Movie Dataset Using Closet Algorithm

Efficient mining frequent closed movie-sets (CLOSET) is used to prevent generating a huge number of movie-sets. CLOSET is an FP-tree-based database projection method for efficient mining of frequent closed movie-sets. This system explores the partition-based projection mechanism for scalable mining. It develops a single prefix path compression technique to identify frequent closed movie-sets quickly [7].

For example, a movie information stored in database is shown in Table 1. Firstly, movies are named with code no (001, 002, 003, 004, 005, 006).

Table 1. Movie information

Code No	MovieName	Actor	Actress
001	The stars of the future	Aung Ye Lin	May Thet Khine
002	Kay Sar	Nay Toe	Moe Hay Ko
003	Nothing is free	Khant Si Thu	Thet Mon Myint
004	Smaller or Bigger	Kyaw Ye Aung	Soe Myat Thu Zar
005	Secretive love II	Zay Ye Htet	Khin Wint War
006	Ko Kyi	Pyay Ti Oo	Yu Thanda r Tin

This system considered the movie transactions database (TDB) for experiment. The movie transactions database is shown in Table 2.

Table 2. The movie transactions database (TDB)

TransId	Movies in Transaction
T1	003, 005, 006,001,004
T2	005,001,002
T3	003, 005,006
T4	003, 006, 001,004
T5	003, 005,006

In the movie transactions database (TDB), the set of frequent item list in support descending order is:

f_list ={003:4, 005:4, 006:4, 001:3, 004:2} By using Divide search space method, all frequent movie-sets can be divided into 5 non-overlap subsets based on f_list:

- The ones containing 004.
- The ones containing 001 but no 004.
- The ones containing 006 but neither 001 nor 004.
- The ones containing 005 but neither 006 and 001 nor 004.
- The ones containing only 003.

Let minimum support count=2.

Then, the conditional database is constructed according to f_list. This finds the closed frequent movie-sets for each frequent item. The following figure 1 is the frequent closed movie-sets for code no (004).

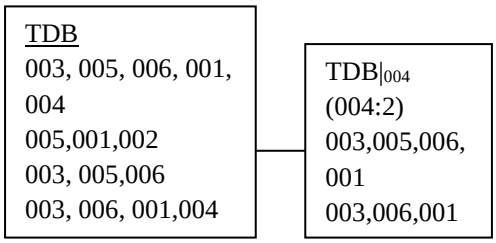

Figure 1. Frequent closed movie-sets for 004

From the above figure 1, F.C.I is {003, 006, 001, 004 : 2}.

In the database, the system finds all transactions containing item 004. And the conditional database is constructed for 004. It is denoted as $TDB|_{004}$. In this database, only transactions containing 004 are included. But item 004 is omitted in each transaction since it appears in every transaction in the $TDB|_{004}$. All transactions containing in TDB|004 are all frequent. Then, the frequency of each item is counted. The support of 004 is 2. Items 003, 006 and 001 appear twice respectively in $TDB|_{004}$. This means that every transaction containing 004 also contains 003, 006 and 001. Therefore, movie-sets {003, 006, 001, 004: 2} is a frequent closed movie-sets. When frequent movie-sets containing 001 is searched, 004 have been found in $TDB|_{004}$ are omitted. And any items have been found are omitted in other transactions.

4.2. Generating Association Rules from Frequent Closed Movie-Set

If the frequent closed movie-sets from transactions in a database TDB have been found, it is straightforward to generate strong association rules. It is considered where strong association rules satisfy both minimum support and minimum confidence. This can be done using the following equation:

Confidence is the measure of the strength of implication. Confidence (A=>B) = P (B|A) = support_count(A U B) / support_count(A). The conditional probability is expressed in terms of movie-sets support count, where: support_count(AUB) is the number of transactions containing the movie-sets AUB. Support_count(A) is the number of transactions containing the movie-sets A. Based on this equation, association rules can be generated as follows:

1. For each closed frequent movie-sets l, generate all nonempty subsets of l.
2. For every nonempty subset s of l, output the rule "s => (l-s)" if support_count(l)/ support_count(s) >= min_conf, where min_conf is the minimum confidence threshold.

For example, a movie transactions database shown in Table 2 is considered. It is supposed that the data contain the closed frequent movie-sets l = {003, 005, 006}. The non empty subsets of l are: {003, 005}, {003, 006}, 005, 006}, {003}, {005} and {006}. The resulting association rules, six rules that satisfy the minimum support count = 2 are shown below. Each listed with its confidence. Confidence from these movie-sets are as follows:

003 and 005=>006 Conf=3/3=100% ----------(1)

003 and 006=>005 Conf=3/4=75% ----------(2)

005 and 006=>003 Conf=3/3=100% ----------(3)

003=>005 and 006 (Conf=3/4=75%) ---------(4)

005=>003 and 006 (Conf=3/4=75%) ---------(5)

006=>003 and 005 (Conf=3/4=75%) ---------(6)

If the minimum confidence threshold is 75%, then rule 1 and rule 3 above are the results of output, because these are the only ones generated that are strong [7]. According to the generated strong rules, users know which movie-sets are really interested.

5. Implementation

This system implemented for generating strong association rules by using CLOSET algorithm. In implementing this system, the movie transaction database (TDB) is used to generate the frequent movie list. Firstly, data about the information of movie borrowing shop is stored into the movie database. In one transaction, movie-sets of movies borrowed by borrowers are contained. Then, transaction of movies is stored in database. This system scan database and compute frequent item list, called f_list. Then, f-list is defined by the user-specified minimum support count and sorted by descending frequency order.

Divide search space methodology is employed supported f_list. And conditional database is constructed from every frequent movie-sets to search out frequent closed movie-sets. The pattern growth is achieved by the concentration of the suffix pattern with the frequent patterns generated from a conditional FP-tree.

From the conditional FP-tree, the closed frequent movie-sets using CLOSET algorithm is found. To facilitate tree traversal, a conditional database is constructed. So every movie-set points to its occurrences within the tree via a sequence of node-links. The tree with the associated node-links is obtained after scanning all of the transactions. During this manner, the manner of movie-sets in database is remodeled to it of mining FP-tree.

Then association rules for these movie-sets are generated. The confidence and correlation for the movie-sets of movies are calculated. Finally, the system generates the association rules that are strong. So, the users know which movies are really interested due to the knowledge of strong association rules. The system flow diagram is described in Figure 2.

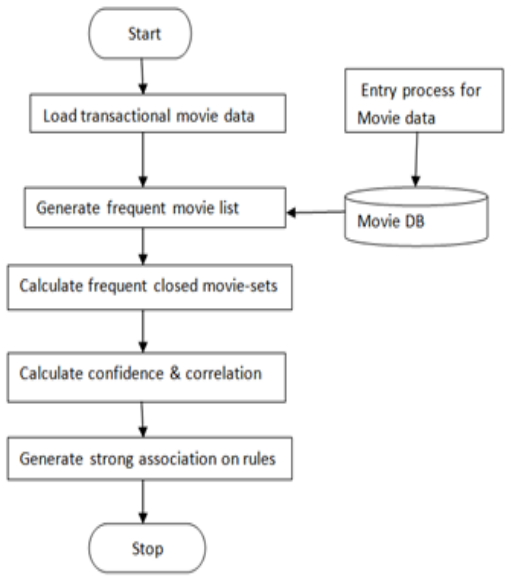

Figure 2. System Flow Diagram

In this system, CLOSET algorithm is used together with divide search space methodology. From the above generated frequent closed movie-sets of movies, their associated information and their confidences are appeared. The set of frequent closed movie-sets with their associated confidences are shown in Table 3.

Table 3. Associated confidence of movie-sets

F.C.I (A)	F.C.I (B)	Confi (%)	MovieInfo
003	005	100	003= Nothing is free (Khant Si Thu)(Thet Mon Myint) 005=Secretive love II (Zay Ye Htet) (Khin Wint War)

003	006	75	003=Nothing is free (Khant Si Thu)(Thet Mon Myint) 006=Ko Kyi (Pyay Ti Oo)(Yu Thandar Tin)
005	006	100	005= Secretive love II (Zay Ye Htet) (Khin Wint War) 006= Ko Kyi (Pyay Ti Oo)(Yu Thandar Tin)
003	005,006	75	003= Nothing is free (Khant Si Thu)(Thet Mon Myint) 005= Secretive love II (Zay Ye Htet) (Khin Wint War) 006=Ko Kyi (Pyay Ti Oo)(Yu Thandar Tin)

5.1. Rule Interestingness Measure by Correlation Analysis

A correlation measure can be used to augment the support-confidence framework for association rules [15]. There are various correlations that measure to determine which would be good for mining large data sets. Lift is a simple correlation measure. The occurrence of movie-sets A is independent of the occurrence of movie-sets B if it is P(A U B) = P(A)P(B). Otherwise, movie-sets A and B are dependent and correlated as events. This definition can easily be extended to more than two movie-sets.

Although minimum support and confidence threshold facilitates the exploration of a good number of uninteresting rules. Several rules generated are still not interesting to the users. This is especially true when mining at low support threshold or mining for long patterns. Even strong interesting rules can be uninteresting or misleading. The support-confidence framework may be supplemented with extra powerfulness measures supported on correlation analysis.

6. Conclusion

CLOSET algorithm discovers interesting association or correlation relationships among huge number of data items. This system generates interesting and strong association rules. And correlation between strong association rules is calculated based on lift method. So, the user can evaluate that which movie-sets are really interesting. The user knows closed frequent movie-sets of movies and which movies are interesting using the CLOSET algorithm. As a result, association rule extracted from the closed frequent movie-sets l= {003, 005, 006} that satisfy the confidence threshold value. So, rule (1) and rule (3) are the results of output. Moreover, this paper is intended to extract association rule mining. The experiment results perform using movie-sets in database and confidence threshold by using CLOSET algorithm. In future, this system can be improved by other association rule mining algorithms.

Acknowledgements

I would like to take this opportunity to express my sincere thanks to all who gave me a lot of valuable advice and information. I am graceful to all respectable people who directly or indirectly helped in the task of developing this paper. Finally, I would like especially to thank my colleagues and friends for the completion of this paper.

References

[1] Ming-Syan Chen, Jiawei Han, P.S.Yu, Data mining: an overview from a database perspective, IEEE Transactions on Knowledge and Data Engineering, Volume:8, Issue 6 ISSN: 1041-4347, 866-883.

[2] Springer, "Principle of Data Mining, Undergraduate Topics in Computer Science".

[3] N. Pasquier, Y. Bastide and R. Taouil et al. (1999). Discovering frequent closed itemsets for association rules. In Proceeding of the 7th Int'l Conference on database theory, Jerusalem, Israel, January, pp.398-416.

[4] C.I. Ezeife and Dan Zhang, "TidFP: Mining Frequent Patterns in Different Databases with Transaction ID", School of Computer Science, http://www.cs.uwindsor.ca/~cezeife.

[5] Qiankun Zhao, Sourav S. Bhowmick, Association Rule Mining:

[6] Dungarwal Jayesh M and Neeru Yadav, "A Review paper for mining Frequent Closed Itemsets", S.V.C.S.E. Alwar Rajasthan –India and Prof S.V.C.S.E Alwar Rajasthan – India.

[7] Jian Pen, Jiawei Han, and Runying Mao, "An efficient Algorithm for mining Frequent Closed Itemsets", Intelligent Database Systems Research Lab, Canada V5A 1S6, {peijian, han, rmao}@cs.sfu.ca.

[8] Jiawei Han and Micheline Kamber, "Frequent Item set Mining Methods, Data Mining– Concepts and Techniques", Chapter 5.2, Julianna Katalin Sipos.

[9] Jiawei Han hanj@cs.uiuc.edu, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", University of Illinois at Urbana-Champaign.

[10] Lai Lai Win, Khin Myat Myat Moe, Computer University (Magway), "Mining Association Rules by using Vertical Data Format", lailaiwin.myn@gmail.com.

[11] Mihir R Patel, Dipak Dabhi, "An Extensive Survey on Association Rule Mining Algorithms", CGPIT, Bardoli, India.

[12] V. Purushothama Raju and G.P. Saradhi Varma, "Mining closed sequential pattern in large database", Department of Information Technology S.R.K.R. Engineering College, Bhimavaram, A.P., India.

[13] Charu C. Aggarwal, Jiawei Han Editors, "Frequent Pattern Mining".

[14] David Hand, Heikki Mannila and Padhraic Smyth, "Principles of Data Mining" ISBN: 026208290xThe MIT Press © 2001 (546 pages).

[15] Kyae Hmon, "Applying Associative Rule Mining And Correlation Analysis On Software CD Selection Data", Computer University (Monywa), Myanmar, mirror315@gmail.com.

User-Based Collaborative Filtering Recommender System for Books

Thidar Nwe
UCS(PKKU)
thidarnwe@ucspkku.edu.mm

Thin Thin Nwe
UCS(PKKU)
thinthinnwe@ucspkku.edu.mm

Tin Tin Thein
UCS(PKKU)
tintinthein@ucspkku.edu.mm

Abstract

Recommender Systems (RSs) can help the user in navigating through information on the internet and provide information that meets the user's satisfaction and needs. With the rapid development of the internet, information and data have exploded in size, and it is more difficult for people to obtain accurate and efficient information in time. RSs analyze user behavior and present products relative to the user's interest. Collaborative Filtering (CF) is a technique used by recommender systems on the web. User-Based collaborative filtering is successful to predict customer behavior and activities which may involve user interests. This paper presents a recommender system for books based on user-based collaborative filtering.

Keywords: Recommender system, Collaborative Filtering, user-based collaborative filtering, Books

1. Introduction

The goal of data mining is to analyze processes and extract knowledge from data in the context of large databases by using different data mining methods and techniques. A large number of data are available in the information industry, all these data are not usable until converted into meaningful or useful information. E-commerce companies have recognized the need to focus on providing a compelling shopping experience. A website using a recommender system can more effectively provide a user with a useful and relevant suggestion that could fulfill his current information requirement. Recommender systems can be implemented in any domain from E-commerce in the form of personalized services. Customers and manufacturers can get benefit from these systems by suggesting items to customers. Recommender systems consist of parts, user, and item. A user may be a customer or consumer of any product or item, who get the suggestions. In the case of a collaborative filtering approach, build the model from various aspects of users upon their past behaviors. These include items purchased by the user previously as well as the rating is given by the users for a particular item. This can be represented by an m-by-n matrix (called a rating matrix) with m refers to customers (rows) and n refers to products (columns). The similarities between the products can be calculated by using item-based filtering or user-based filtering.

2. Related Works

Anisha Saehan and Vineet Richa Riya are given a "Survey on Recommender System based on Collaborative Technique" that helps to discover the information of the user's interest. They recommend the books using four types of filtering technique, which includes demographic technique, content-based filtering, collaborative filtering, and hybrid method.[1]

Prem Melville and Vikas Sindh ani are introducing a paper on "Recommender System" that defines different recommendation methods and approaches. They also tried to define the common challenges and limitations in the recommendation system. [2]

Chavis Rana and Sanjay Kumar Jain proposed a paper on "Building a Book Recommender system using time-based content

filtering". They expressed that, recommendation systems a new generation tool for helping the people in navigating information through the internet and retrieving information according to their preferences. At their work, they used a content-based approach with a new dimension called the temporal dimension. With the help of a counter each time the item gets an update with the passage of time. [3].

M.s Pooja Malhotra, Sossaman Ragpicker, and Ms. Darshana Bhatt proposed a paper on "Book Recommendation System". They introduce a new approach for recommending books to the buyers by considered many parameters like the content of the book and the quality of the book. [4]

Ms. Para veena Mathew, Ms. Binky Kuriakose and Mr. Vinayak Hegde proposed a paper on "Book Recommendation System through Content-Based and Collaborative Filtering Method". The authors proposed a system that saves details of books purchased by the user. From these Book contents and ratings, a hybrid algorithm using collaborative filtering, content-based filtering, and association rule generates book recommendations. [5]

Akon, E.U., Eke, B.O. and Agba, P.O presented "An improved online book recommender system using collaborative filtering algorithm". They proposed a model that generates recommendations to buyers, through an enhanced CF algorithm, a quick sort algorithm, and Object-Oriented Analysis and Design Methodology. Scalability was ensured through the implementation of Firebase SQL. This system performed well on the evaluation metrics. [6]

2.1. Recommender Systems

Recommender systems are also known as information filtering systems that suggest items according to the user's preferences. Recommender systems can be applied in many areas such as books, movies, and music. These systems can be used content-based filtering, collaborative filtering, and hybrid approach.

2.1.1. Content-based filtering

The content-based filtering approach tries to build a model, based on the available "relations", which express the observed user-item interactions. As an example, different ages of people want to read different kinds of books such as classics, tragedy, science fiction, or humor. And there may be a different sex. These relations become the attributes of the products. By using these attributes, similarities can be derived between the products. The advantage of content-based filtering is the possibility of precisely defining relations between products. But on the other hand, this approach requires the manual definition of a great number of additional information, e.g., keywords and attributes for each product. And also uses complicated data mining techniques to generate recommendations.

2.1.2. Collaborative filtering

The Collaborative filtering approach needs information based on past interactions recorded between users and items in order to produce new recommendations. These interactions are stored in the "user-item interactions matrix" or called "rating-matrix".

These past user-item interactions are used to detect similar users and/or similar items and make predictions based on this estimated nearness. Collaborative filtering is divided into two sub-categories that are generally called model-based and memory-based approaches. Model-based approaches assume an underlying "generative" model that explains the user-item interactions and try to discover it to make new predictions. Memory-based approaches directly work with values of recorded interactions and are essentially based on nearest neighbors search (for example, find the closest users from a user of interest and suggest the most popular items among these neighbors). This approach can be subdivided into two ways: user-based approach and item-based approach.

User-based approach: users with the same choices form a neighborhood. If an item is not rated by the user, but it has been rated by other users of the neighborhood, then it can be

endorsed to the user. Hence the user's choice can be predicted based on the neighborhood of similar users. This paper used the user-based collaborative filtering method.

Item-based approach: similarity between the group of items rated emphatically by the user and the required item is calculated. The items which are very similar are selected. The Recommendation is computed by the weighted means of the user's ratings of the same item.

2.1.3. Hybrid approaches

To obtain accurate and faster recommendations, different recommender systems use hybrid approaches. Some hybrid approaches are:

- Separate execution of producers and connecting the results
- Using some content filtering guidelines with community collaborative filtering
- Using some principles of collaborative filtering in content filtering recommender
- Using both content and collaborative filtering in a recommender

3. Proposed Work

The purpose of this user-based recommender system is to recommend books to the user that meets the user's preference. This recommender system used a user-based collaborative filtering approach. This approach attempts to predict the products based on the behavior and activities of other users related to the customer. The recommender system is expressed as a Block diagram in Figure 1.

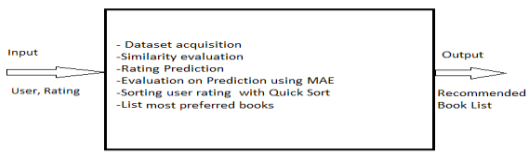

Figure 1. Block diagram of proposed system

The system used user information and user ratings for books, which they choose as input, and the system will combine with pre-recorded data and calculate the predictions using the cosine similarity method to suggest the books to the user. The accuracy of prediction can be evaluated by using the Mean Absolute Error (MAE) formula. Before the system displays the resulted books to the user, these books are sorted in descending order based on the user's rating by using a quick sort algorithm. Then the ten most preferred books will be displayed to the user.

The proposed system architecture is described in Figure. 2.The elements in the system are:

User: The user represents the individual or client that utilizes the interface. User activities include querying the web application for specific choices, viewing of reading materials (books), adding books to cart, completing the purchases or rental processes and manually rating books, etc.

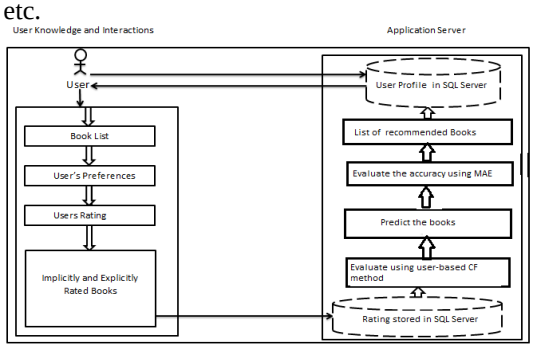

Figure 2. Architecture of the proposed system

Book List: A book list in the proposed system refers to a collection of recommended or non-recommended books presented to the user through the web interface.

User's Preferences: The user's preference is used to describe the user's preferred choice of suggested books. A collection of user choices can be used to suggest the effective recommendation of books for other users.

Users rating: The user rating describes the act of the user providing ratings for the books in the system.

Implicitly and Explicitly Rated Books: This represents a collection of book ratings which are obtained by explicitly providing users with an

interface/option to rate books online. Rated books data is also obtained by implicitly observing and taking note of how a user interacts with the books.

Rating stored in SQL Server: Different rated books and user lists are stored into a SQL Server database and used to perform the recommendation of books.

User-based Collaborative Filtering: This involves the generation of book recommendation lists for users using the inputs in the database. The recommender system performs the computation of similarity among items with other rated items in the database for a given user using cosine similarity model given as:

$$sim(x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{||\vec{x}||_2 \times ||\vec{y}||_2} = \frac{\sum_{s\in S_{xy}} r_{x,s} r_{y,s}}{\sqrt{\sum_{s\in S_{xy}} r_{x,s}^2}\sqrt{\sum_{s\in S_{xy}} r_{y,s}^2}}, \quad (3.1)$$

x, y = users

$r_{x,s}$ = rates of user x on item s.

$r_{y,s}$ = rates of user y on item s.

After the similarities have been calculated, system predict the recommended values for the user by using the equation

$$r_{c,s} = k \sum_{c' \in C} sim(c,c') \times r_{c',s}$$

c = user to rate on item s.

c' = other users.

After the recommended results are obtained. These results are evaluated by MAE to check the accuracy of the prediction.

$$\text{MAE} = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \quad (3.2)$$

Where

N = the total number of actual ratings in an item set

p_i = the prediction of user's ratings

q_i = corresponding real ratings data set of users.

The lower the MAE value means the better the prediction of the recommender system.

List of Recommend Books: Quicksort algorithm is applied to sort the rated items. And finally, the results are displayed to the user and forwarded into the user profile database.

QUICKSORT Algorithm

Procedure QUICKSORT(S);

1.If S contains at most one element then return S.

else

2. begin Chosen an element **a** randomly from S;

3.Let S1, S2 and S3 be the sequence of element in S less than, equal to, and greater than **a**, respectively;

4.return (QUICKSORT (S1) followed by S2 followed by QUICKSORT (S3))

End.

User Profile: The user profile describes a database containing a user's personal details, their implicit/explicit ratings for books, and a list of recommendations provided using user's similarity with other users.

4. Implementation

The system is implemented by using C#.net as front end and SQL Server as backend on Intel Core2duo processor with 2GB RAM and 300 GB hard disk.

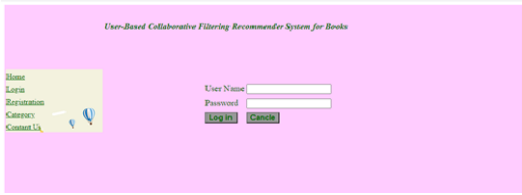

Figure 3. Login Page

If the user has already registered, the user can log in to the system by using a user name and password. If the user is not registered, the user has to register into the system by fill out the registration form.

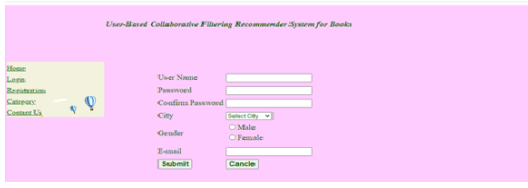

Figure 4. Registration Page

After filling a user name and password, the system shows the most rated books and the recommended books.

Figure 5. Recommendation Lists

5. Experiment Results

In this section, we describe the books dataset. The Collaborative filtering method will predict the rating value of the items in the dataset that the user did not rate. That predicts rating values accuracy are measured by using MAE.

5.1. Dataset

The dataset of books is collected from Burma-Myanmar site (https://www.goodreads.com/ shelf / show/burma-myanmar). In this paper, there are 10 unknown value of rating on books as a sample data set, extracted from the used users-items metrices (100*100) size. They are evaluated by cosine similarity method in next section 5.2.

Table 1. Data set

	B1	B2	B3	.	B5	.	B10	.	B14	.	B16	B17	.	B100
U1	4	4	?		3		5		5		4	3		4
U2	5	?	3		2		3		3		5	2		5
U3	4	5	?		5		4		4		4	5		5
U4	?	4	5		3		5		5		3	5		4
:														
U14	?	1	3		5		5		5		3	5		1
:														
U41	3	3	4		5		4		4		3	?		3
..														
U	3	5	5		?		4		5		3	3		5
59														
:														
U70	3	5	5		4		5		?		3	4		5
:														
U80	5	5	5		4		5		5		?	4		5
:														
U99	5	4	2		?		2		2		5	5		4
U100	2	3	4		5		4		4		2	5		3

B1, B2, …, B100 = Books
U1, U2, …, U100 = Users

5.2. Predicted Rating Value Evaluation

In this section, the user's rating values are predicted by the use of a collaborative filtering method based on table 1. MAE (Mean Absolute Error) is also used to measure the variation between the predicted value and the actual user rating value. In our system, how truth the predicted value is on how many left the MAE value behinds of the decimal. As a result, to predicted values varied from MAE are shown in Table 2.

Table 4. Experiment Result

User id	Book id	Predict Value	Accuracy using MAE
U1	B3	3.8493	0.0493
U2	B2	3.9963	0.0563
U3	B3	3.8060	0.0060
U1	B4	3.8448	0.0048
U14	B1	3.7741	0.0341
U41	B17	3.9300	0.0700
U59	B5	3.9554	0.0354
U70	B14	3.8422	0.0021
U80	B16	3.7023	0.0023
U99	B5	3.9468	0.0268

6. Conclusion

Recommendation system is widely used from the last decades. This system will record the details of the books that users have previously purchased and search for books by user history. The proposed system is to collect the buyer's interest and recommends the book for other users. This recommendation system also uses a collaborative filtering method to give a stronger recommendation. It used a real-time database, an efficient quicksort algorithm, and a cosine similarity algorithm to improve on recommender system.

7. Future Work

This recommender system depends on the ratings given by users. Thus, trust is a major issue, the feedback and ratings given by users cannot be proved as these are actual or not. So, the system does not solve the trust issue. Therefore, future research should focus on resolving this issue.

References

[1] Aisha Saehan and Vineet Richa Riya, "Survey on Recommender system based on Collaborative Technique", Department of Computer Science And Engineering, International Journal of Innovations in Engineering and technology (IJIET), vol-2, pp 1-7, April 2013.

[2] Prem Mlville and Vikas Sindh ani, "Recommender System", IBM T.J. Watson Research Centre Yorktown Heights, pp. 1-18.

[3] R.ana Chav and Jain Sanjaya Kumar, "Building a Book Recommender system using time based content Filtering", University Institute of Engineering and Technology, vol. 11, no. 2, pp. 2224-2872, February 2012.

[4] M.s PoojaMlhotra, Sossaman Ragpicker and MS. Darshana Bhat, "Book Recommendation System", International Journal for Innovative Research in Science & Technology, Volume 1, issue 11, April 2015,ISSN 2349-6010

[5] Mathew, P., Kuriakose, B. And Hegde, V. (2016). "Book Recommendation System through content based and collaborative filtering method." Proceedings of International Conference on Data Mining and Advanced Computing (SAPIENCE)

[6] Akon, E.U., Eke, B.O. and AS agba, P.O., 2018. "An improved online book recommender system using collaborative filtering algorithm." International Journal of Computer Applications (0975- 8887) Volume 179-No.46, June 2018.

Performance Evaluation of Frequent Pattern Mining (Apriori and FP-Growth)

The` Su Moe
UCS(PKKU)
thesumoe1@gmail.com

Cho Cho Khaing
UCS(PKKU)
chokhaing28@gmail.com

Zin Mar Shwe
UCS(PKKU)
zinmarshwe1976@gmail.com

Abstract

Frequent pattern mining is an area of extensive research in data mining because it is important in many practical applications. Many algorithms are used to mine frequent patterns with different performance on different datasets. The system uses web-usage mining technology to generate frequent patterns that are collected from web server log files. Web usage mining can be further classified according to the type of usage data considered. User logs for web server data are collected by the web server and usually include IP address, page browsing, and access time. The system uses the Apriori and FP-Growth algorithms to find frequent patterns. The system compares the evaluation results of these two algorithms based on time efficiency and memory usage.

Keywords: Frequent pattern mining, web usage mining, web logs, Apriori, FP-Growth, time efficiency

1. Introduction

For many years, frequent pattern mining has been an important topic in data mining. Significant progress has been made in this field, and many effective algorithms have been designed to search for frequent patterns in transactional databases. Agrawal et al. (1993) firstly proposed a pattern mining concept in form of market-based analysis for finding the association between items bought in a market. This concept uses transactional databases and other data repositories to extract any structure of interest, interesting relationships, or a set of frequent patterns [1].

Frequent pattern mining (FPM) is the most important field of association rule mining. It was originally developed for market basket analysis. FPM is a fundamental phenomenon and plays an important role in many applications [2] [3]. FPM can be tested in separate data formats, such as transactional databases, sequence databases, streams, strings, spatial data, and graphs [4] [5]. Most surveys of frequent itemset mining to focus on the performance differences between itemset algorithms that occur frequently in a single dataset [6].

As the amount of data available on the network explodes, it is a real need to discover and analyze useful information from the Web. Web-used mining is an application of data mining techniques to discover interesting usage patterns from Web usage data and to better understand and meet web-based needs. Web-used mining has three main phases: data preprocessing, pattern detection, and pattern analysis. The data collected by the Web server includes the user's IP address, page reference, and access time, which is the primary input. In this paper, the performance of Apriori and FP-Growth algorithm using web log data is tested.

2. Related Works

In this study, many technological experiments are conducted using the mining association's rules [7] [8]. In the field of correlation rule mining, the Apriori algorithm is the most widely used algorithm for generating candidate patterns [2]. This is a sensible search of the level. It digs

up patterns often through countless database scans. Based on the Priori algorithm, several improvements or adjustments have also been made on many algorithms, such as the AprioriTid algorithm [2]. It restores more time but memory usage is minimal.

UT-Miner [9] is a special temporary Apriori algorithm for sparse data. In sparse data, most transactions are different from each other. Array structures are used to improve mining performance because it is based on the Apriori algorithm, even if it does not provide a guarantee of run-time and memory usage.

Another achievement of frequent pattern mining is the FP-Growth algorithm [10]. A high-energy algorithm called FP-Growth is introduced to establish a frequent pattern tree structure called the FP tree to overcome two flaws in the Apriori algorithm. First, no candidate patterns are detected. Second, the database scan is performed only twice. It takes a way called "divide and conquer". Improved frequent pattern (IFP) growth techniques [12] have been proposed for detecting frequent patterns. This algorithm uses low memory and shows improved testing effectiveness based on the FP tree algorithm.

3. Frequent Pattern Mining

Frequent patterns are patterns that are frequently displayed in datasets, such as itemset, subsequence, and substructures. For example, a set of items that are frequently displayed in a transaction dataset, such as milk or bread, is ones that occur frequently. Itemset mining was introduced frequently as a method of frequently grouping items in the basket/transactional databases that contain these items [13]. A database consists of a series of baskets that resemble customer orders. These orders consist of a single basket of numbers of items. Companies such as Amazon, Netflix, and other online retailers frequently use product sets to recommend the purchase of other products based on consumers' past purchase history and history of other products like baskets [14]. The following data shows the baskets that can be used for frequent itemset mining, and each row represents a single item set.

[mp3player usb−chargerr book−dct book−ths]
[mp3player usb−charger]
[usb−charger mp3player book−dct book−ths]
[usb−charger]
[book−dct book−ths]

From the above baskets several frequent itemset can be defined. These are sets of items that frequently occur together, some of which are:
[mp3player USB-charger]
[book-dct book-ths]
…

A simple visual analysis of the data shows that mp3players and USB-chargers often happen at the same time. Likewise, book-dct and book-ths often happen together. The itemset algorithm, which occurs frequently, uses a variety of statistics to find which itemset are included.

3.1. Apriori Algorithm

Apriori is the first algorithm to mine frequent patterns. It was written by R Agrawal and R Srikant in 1994. This algorithm is suitable for databases based on a horizontal layout. It is based on Boolean association rules that use build and test methods. It uses BFS (breadth-first search). Apriori uses frequently k-item sets to find larger k+1 itemset. In the Apriori support count for each project, the algorithm scans firstly the database for all frequent items based on support. The frequency of an item can be calculated by calculating the frequency of the items that occur in all transactions [16]. All infrequent items are deleted.

Apriori property: In all subsets of frequent, non-empty itemset also occur frequently. Apriori follows a two-step approach: the first step is to add two item sets that contain the K-1to Kth pass common items. The first channel starts with a single item, and the result set is called the candidate set Ck. In the second step, the algorithm counts the number of occurrences of each candidate set and crops all infrequent itemset. If no further extensions are found, the algorithm terminates. The former algorithm is shown in Figure 1.

C_k: Candidate itemset of size k

L_k : frequent itemset of size k

L_1 : {frequent item}

for (k=1; L_k!=∅ ;k++) do

C_{k+1}=candidates generated from L_k;

for each transaction t in database do

increment the count of all candidates in C_{k+1} that are contained in t

L_{k+1}= candidates in C_{k+1} with min_support

end

Return $\cup_k L_k$

$$\text{Confidence(A=>B)} = \text{P(B/A)} = \frac{Support\ Count(A \cup B)}{Support\ Count(A)}$$

Figure 1. Apriori algorithm

3.2. FP-Growth Algorithm

The idea of a tree-based algorithm that mines frequent patterns in the database is marked as FP growth (Han et al. 2000) [18]. It is suitable for projection type databases. [17] It uses the method of division and conquest. If it is not needed frequent project set suggestions, patterns should fairly often be done from FP tree mining. The first step generates a list of frequently occurring itemset and sorts them in the order of decrement support. This list is represented by a structure called a node. Each node (root node) in the FP tree contains the item name, the support count, and a pointer [16] that points to the node with the same item name in the tree. These nodes are used to create the FP tree. Common prefixes can be shared during the construction of the FP tree. The path from the root node to the leaf node is ordered in the supported non-growth order. After the FP tree has been built, frequent patterns are extracted from the LEAF tree in the leaf node. Each prefix path subtree is processed recursively to mine frequently occurring sets of items.

Procedure FP Growth (Tree, α)

if Tree contains a single path P then

  for each β = comb.of nodes in P do

    pattern = $\beta \cup \alpha$

    sup = min(sup of the nodes in β)

else

    for each a_i in the header of Tree do

    generate pattern= $\beta \cup \alpha$

    sup=a_i.support

    construct $\beta's\ conditional\ pattern\ base$

FP Tree= construct $\beta's\ conditional$ FP Tree

If FP Tree!= 0 then

FP Growth (FP Tree β)

Figure 2. FP-Growth algorithm

The projection layout allows FP-Growth to consume minimal memory and improve storage efficiency. FP tree variations are conditional FP trees that are built by taking into account transactions that contain a specific item set and removing them from all transactions. Another variation is the parallel FP extension (PFP), which is recommended to parallelize the FP tree on distributed computers [19]. FP growth has been improved using the prefix tree structure by Grahne and Zhu [20]. The FP-Growth algorithm is as shown in Figure 2.

4. Proposed System Design

The proposed system flow is as shown in Figure 3.

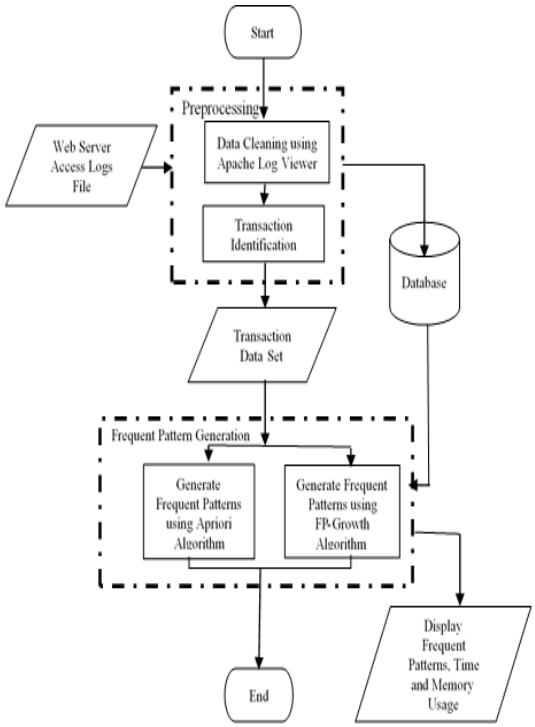

Figure 3. System flow diagram

The system uses the access Web Log as input. Next, necessary data are taken from the access log file called the preprocessing step, as shown in Figure 4. When the preprocessing step is completed, the Apriori and FP-Growth algorithms are used to generate frequent patterns and processing times.

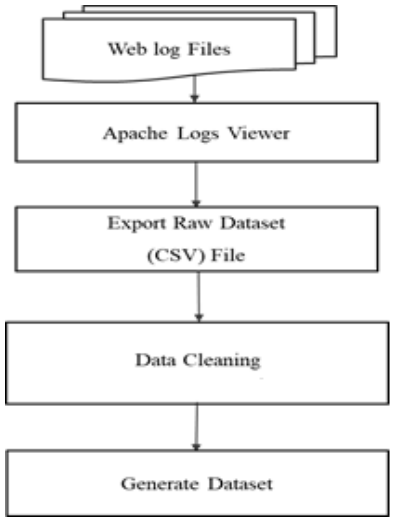

Figure 4. Preprocessing Step

5. Experimental Results

In this paper, five Web Log datasets (kosarak, accident, mushrooms, retail, record link) were tested, and these two types of experiments (processing time and memory usage) were conducted. Table 1 shows the total transactions for the five datasets.

Table 1. Number of transactions for each dataset

Datasets	Number of Transactions
Kosarak	990002
Accidents	340183
mushrooms	8416
Retail	88162
RecordLink	574914

Table 2.Time (milliseconds) comparison of Apriori and FP-Growth

Datasets	Apriori	FP-Growth
kosarak	1924	1909
accidents	9684	3989
mushrooms	461	127
retail	235	221
RecordLink	2142	1428

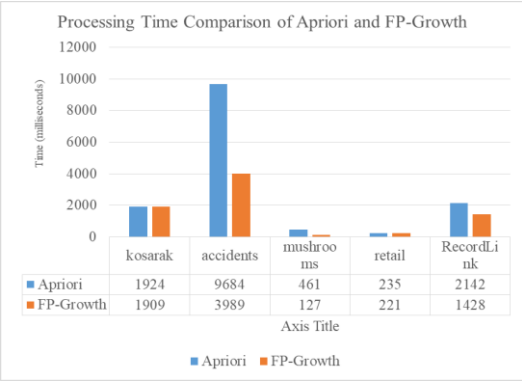

Figure 5. Experimental result I

According to the experiment I as shown in Table 2 and Figure 5, the processing time of Apriori was 461 milliseconds and 127 milliseconds in the FP-Growth algorithm to produce frequent patterns from the mushroom dataset. In accident datasets, the Apriori was 9684 milliseconds and the FP-Growth was 3989 milliseconds. So, the FP-Growth algorithm is almost twice as fast as the Apriori in the production of frequent patterns.

Table 3. Memory usage (megabits) comparison of Apriori and FP-Growth

Datasets	Apriori	FP-Growth
kosarak	125.33	250.5
accidents	181.69	241
mushrooms	22.32	39.42
retail	77.89	191.4
RecordLink	258.81	270.39

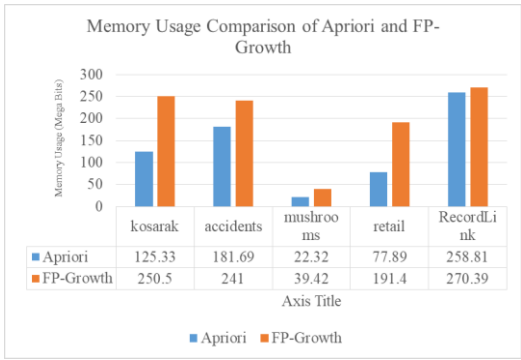

Figure 6. Experimental result II

In experiment II, as shown in Table 2 and Figure 6, Apriori's memory usage was 22.32 megabits, and the memory usage in the FP-Growth algorithm was 39.42 megabits to generate frequent patterns of mushroom datasets. In the accident's dataset, Apriori was 181.69 megabits and FP-Growth to 241 megabits. As a result, the memory usage of the FP-Growth algorithm is greater than Apriori in the generation of frequent patterns.

6. Conclusion

Frequent pattern mining is the most important step in related rules and ultimately useful in many applications, including market basket analysis, clustering, series analysis, games, decision-making, object mining, and site navigation. In this paper, the pattern mining algorithms, i.e. "Apriori" and "FP-Growth" are evaluated. The Apriori algorithm is an iterative method called level search. This algorithm uses the breadth-first search (BFS) and candidates' itemset is required to produce item generations. It undergoes the repeated scans of the data. On the other hand, the FP-Growth algorithm adopts a "divide and conquer" approach that does not require the generation of candidate sets. The data scan is not repeated. Experimental results were evaluated based on performance characteristics such as runtime and memory usage. Therefore, the FP-Growth algorithm was found to have better performance than the Apriori algorithm

while the Apriori algorithm handles less memory usage than FP-Growth.

References

[1] Sourav S. Bhowmick Qiankun Zhao, "Association Rule Mining: A Survey," Nanyang Technological University,Singapore, 2003.

[2] Agrawal, R. and R. Srikant, 1994, "Fast algorithms for mining association rules", Proceedings of the 20th Very Large Databases Conference (VLDB'94), Santiago de Chile, Chile.

[3] Zhihong Deng Zhonghui Wang, 2010, "A New Fast Vertical Method for Mining Frequent Patterns, International Journal of Computational Intelligence Systems", 3(6): 733-744.

[4] Syed Khairuzzaman Tanbeer, Chowdhury Farhan Ahmed, Byeong-Soo Jeong and Young-Koo Lee, 2008, "Efficient single-pass frequent pattern mining using a prefix-tree", Information Sciences, Elsevier ltd, 179: 559-583.

[5] Quang-Huy Duong, Bo Liao, Philippe Fournier-Viger and Thu-Lan Dam, 2016, "An efficient algorithm for mining the top- *k* high utility itemsets, using novel threshold raising and pruning strategies", Knowledge-Based Systems, Elsevier Ltd., pp: 1-17.

[6] D. Burdick, M. Calimlim, and J. Gehrke, "Mafia: A maximal frequent itemset algorithm for transactional databases," in Proceedings of the 17th International Conference on Data Engineering, 2001. IEEE, 2001, pp.443–452.

[7] Imielienskin, T., A. Swami and R. Agrawal, 1993, "Mining Association Rules Between set of items in large databases", in Management of Data, pp: 9.

[8] Mannila, H., R. Srikant, H. Toivonen, A. Inkeri and R. Agrawal, 1996, "Fast Discovery of Association Rules in Advances in Knowledge Discovery and Data Mining", pp: 307-328.

[9] Hamada, M., K. Tsuda, T. Kudo, T. Kin and K. Asai, 2006, "Mining frequent stem patterns from unaligned RNA sequences, Bioinformatics", 22(20): 2480-2487.

[10] Han, J., J. Pei and Y. Yin, 2000, "Mining frequent patterns without candidate generation", Proceedings 2000 ACM-SIGMOD International Conference on Management of Data (SIGMOD' 00), Dallas, TX, USA.

[11] Sourav S. Bhowmick Qiankun Zhao, 2003. Association Rule Mining: A Survey, Nanyang Technological University, Singapore.

[12] Ke-Chung in, I-En Liao, Zhi-Sheng Chen, 2011, "An improved requent pattern growth method for mining association rules", Expert Systems with Applications, pp: 5154-5161.

[13] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD Record, vol. 22, no. 2. ACM, 1993, pp. 207–216.

[14] J. Leskovec, A. Rajaraman, and J. D. Ullman, Mining of massive datasets. Cambridge University Press, 2014.

[15] Goswami D.N. "An Algorithm for Frequent Pattern Mining Based On Apriori", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, Pp. 942-947.

[16] Rahul Mishra, "Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data.", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (4), 2012, Pp. 4662 – 4665.

[17] SathishKumar, "Efficient Tree Based Distributed Data Mining Algorithms for mining Frequent Patterns", International Journal of Computer Applications (0975 –8887) Volume 10– No.1, November 2010.

[18] Han, J., Pei, J., and Yin, Y. 2000, "Mining frequent patterns without candidate generation.", In Proc. 2000 ACMSIGMOD Int. Conf. Management of Data.

[19] Haoyuan Li,Yi Wang,Dong Zhang, Ming Zhang,Edward Chang 2008."Pfp: parallel fp-growth for query recommendation Proceedings of the 2008 ACM conference on Recommender systems Pp. 107-114.

[20] G. Grahne and J. Zhu, May 2003, "High performance mining of maximal frequent itemsets", In SIAM'03 Workshop on High Performance Data Mining: Pervasive and Data Stream Mining.

Country Based Analysis: Relationship between HEXACO Personality Traits and Emoji Use

Yi Yi Win
University of Computer Studies, Meiktila
yiyipku941@gmail.com

Tin Tin Thein
University of Computer Studies, Pakokku
tintinthein@ucspkku.edu.mm

Myat Thet Nyo
University of Computer Studies, Meiktila
myathetnyo05@gmail.com

Wai Wai Khaing
University of Computer Studies, Sittway
waiwaikhine728@gmail.com

Abstract

Communication is a process of conveying a message or information from one person to another, which is very predominant and significant in human existence. The objective of this paper is to estimate user personality traits based on emoji. Two surveys conducted for this study; 65 face type emoji included questionnaire as an emoji measurement tool and HEXACO-PI-R 100 question personality test has been used, as a personality measurement tool in this study. Based on the survey results, final aim to estimate the relationship between user personality traits and emoji by using Pearson Correlation Coefficient and analysis of variance.

Keywords: Pearson Correlation, HEXACO, non-verbal communication

1. Introduction

On the purpose of personality perception according to emotional icons and online communication, there is an influence of perceived receiver personality on one's emoticon usage; because people act and respond according to the personality judgment of the interaction partner in interpersonal conversations [1]. Moreover, human factors such as interpersonal personality perception and relationship may influence the emoticon usage in online communication [2].

Most recently, [4] evaluated the link between emoji usage and human personality in two companion studies. Firstly, they examined psychological factors with utilization of emoticons, including personality factor and secondly they evaluated the accuracy of personality perceptions of facebook users. They revealed that, 'happy face' of emoticons have a positive correlation with Big five personality traits.

There is a very latest study conducted by [3] to examine the utilization of emoji instead of lexical items for assess human personality. The study utilized Big five personality inventory and results could not reveal any association with conscientiousness and openness to experience personality traits between emoji.

To our knowledge, there have been no studies identifying HEXACO six personality traits association with emoji identification. This study addressed the issue of user personality identification according to HEXACO personality traits based on emoji.

2. Proposed System

In this study, user personality assessed by using, one of the prominent non-verbal component of CMC called as emoji. For that purpose, conducted two surveys; emoji survey and personality traits inventory for the sample of 60 participants in five different countries, Thailand, Australia, Myanmar, Sri Lanka, and

Germany. Emoji survey included 65 universal standards (Unicode/ISO 10646) basic face type emoji and asked to score participants' self-identification with emoji by using 5 point Likert scale. The reason to select face type emoji is the highest global emoji usage. According to SwiftKey emoji report in 2015, nearly 59% of emoji usage was face type emoji category

2.1. Online Communication / Computer Mediated Communication (CMC)

In the modern world, Information and communication technology (ICT) has provided a wide space for online communication and the majority have fully controlled by the internet as a network society. Society interacts via online than in person. Figure 1 shows, internet usage per 100 people between the years 2001-2014 [5]. There is a rapid increase of internet usage in both developed countries and developing countries. However, the developed countries still lead figures where their internet usage per 100 people is more than double of the developing countries.

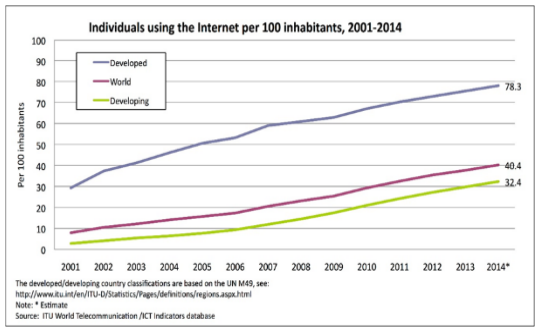

Figure1. Internet Usage (ITU World Telecommunication/ICT Indicators database)

2.2. Emoji Standards

Unicode originated and it is consistent with the ISO/IEC 10646 standard; and uses a unique numerical code across all platforms to identify each and every letter, digit or symbol (Oxford English Dictionary) [6]. Emoji have been present in the Unicode-ISO/IEC 10646 standard for some time now, with the first Unicode characters explicitly intended as emoji added to Unicode 5.2 in 2009.

Figure 2. Unicode Full Emoji Chart (a selected section)
(©Unicode http://unicode.org/emoji/charts/full-emoji-list.html)

Figure 2 illustrates a selected set of Unicode full emoji data on various vendors but with similar coding. For this study, we selected a sample of 65 Unicode-ISO/IEC 10646 emoji, for the survey questionnaire within this full emoji data set and as previously mentioned focus on basic face type emoji.

2.3. Personality Trait

Personality has described as a combination of emotions, attitudes, and behavioral response patterns of individuals. Understanding the philosophy behind "personality" is one of the major research topic for decades and it is described in many ways by the researchers. In this study, HEXACO model for this attempt of identifying user personality traits from their emoji perception according to individuals' personality noting emoji, has used wide across different cultures. HEXACO model, which is introduced by Ashton and Lee is much effective and consistent when discuss cross-cultural traits than the "Big 5 model". Table 1 describes the HEXACO personality traits. HEXACO personality inventory, directly utilized with participants on the website; http://hexaco.org/hexaco-online and after

completing the HEXACO test, participants were asked to send back the downloaded result pdf. Since it has used HEXACO-PI-R authors original website (http://hexaco.org/hexaco-online)

Table 1. HEXACO Personality Traits (Ashton & Lee., 2007)

Factor Name	HEXACO-PI Attributes
Honesty-Humility - H	Sincerity, Fairness, Greed-Avoidance, Modesty
Emotionality - E	Fearfulness, Anxiety, Dependence, Sentimentality
Extraversion - X	Expressiveness, Social Boldness, Sociability, Liveliness
Agreeableness - A	Forgiveness, Gentleness, Flexibility, Patience
Conscientiousness - C	Organization, Diligence, Perfectionism, Prudence
Openness to Experience - O	Aesthetic Appreciation, Inquisitiveness, Creativity, Unconventionality

3. Methodology

This section describe the related methodology of Pearson correlation coefficient which is used to analyze the relationship between HEXACO personality traits and emoji usage. Which are discussed in the following section.

3.1. Pearson Correlation Coefficient

As mentioned above major objective of this study is to ascertain the relationship between the personality traits and the emoji usage in the online communication. Since this study mainly on deriving the relationships between variables, it has conducted correlation analysis on the variables under concern.

Pearson correlation analysis conducted based on the following formula [7-8]:

$$r = \frac{\mathrm{n}(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where,

x = emoji (65 face type of emoji) and y = HEXACO (six type of personality traits).

The Pearson correlation gives an indication on the strength of the relationship between the two random variables x and y. The sign of the correlation coefficient is positive if the variables are directly related and negative if they are inversely related.

If rxy = 0, then x and y are said to be uncorrelated. The closer the value is to1, which means that strong correlation.

In order to make inferences on the population based on the sample, it has conducted a hypothesis testing procedure on the ascertained correlation coefficients.

Hypothesis testing has conducted at the 5% level of significance on the following hypothesis. Country based data are used to analyzed between personality traits and emoji usage in non-verbal communication.

4. Results and Discussion

Used a stratified random sampling method; participants (N = 60) were selected from five different countries; Australia, Germany, Thailand, Myanmar and Sri Lanka. Received 59 responses out of 60 for the online questionnaires; HEXACO-PI-R 100 question personality trait inventory and 65 selected face type emoji for emoji questionnaire. 13 responses were rejected because of incompleteness of responds. After some data screening, 46 responses had selected as valid data. 10 from each Australia, Myanmar, Sri Lanka, and 8 from each Germany and Thailand. The valid data group formed of 32 (69.5%) Females and 14 (30.4%) Males. Participants had contributed voluntarily for the study.

To test the hypothesis that highlight possible correlations between personality traits measured with HEXACO model and usage of emoji in online communication. In this analysis,

participants' attitude towards a sense of emoji usage, personality traits, their demographics and different countries based.

4.1. Country based emoji analysis

This study to examine the HEXACO traits in relation to cross-cultural adjustment or cultural intelligence in five different countries, based on the usage of emoji data in online communication. The same traits have found in every culture, intercultural comparisons and correlations are possible. Although the usage of emoji and the selected group of emoji set are diverse within the analysis of the relationship between personality traits and emoji.

4.1.1. Germany

In the analysis of German participants, they have the five personality traits except "Emotionality", "Honesty-Humility" has a significant relationship with 5 emoji in Table: 2, for example: 🤔 "Thinking face" has a strong correlation 0.865 ($p<0.01$). This table describe the some emoji results which have significant relationship with personality traits. Honesty-Humility represents the tendency to be fair and genuine in dealing with others (St. Catharines, 2007), one of the objective of using HEXACO model is that have included this personality trait and mostly suitable to measure the behaviors across a variety of countries.

4.1.2. Thailand

The results reported in Table: 3 demonstrate the significant relation of HEXACO model and selected emoji set for Thailand. "Honesty-Humility" and "Conscientiousness" personality traits are consequently relevant with 6 emoji. Explained that conscientious people can be good at planning, organizing and have good time management. In the attribute of "Conscientiousness," personality trait has (diligence, organization, perfectionism and prude) that are relevant to emoji usage. 😔 "Pensive face" is one of the reliable significant correlation (-.731, $p<0.05$) for "Conscientiousness" personality trait. This table describe the some emoji results which have significant relationship with personality traits. Other emoji are omitted based on the correlation results with have weak correlation with personality traits.

4.1.3. Myanmar

Pearson correlation analysis for Myanmar people that express in Table: 4. This table describe the some emoji results which have significant relationship with personality traits. Other emoji are omitted based on the correlation results with have weak correlation with personality traits. As per the findings in this study indicate that three aspects of personality traits are related with emoji usage in online communication. The variable with the highest effect was "Conscientiousness" followed by "Emotionality" personality. 7 emoji of explanatory variables have significant effects on the "Emotionality" personality trait. That indicate these 7 emoji (😂 face with a tear of joy, 😀 grinning face, etc…) were able to explain the "Emotionality" personality traits.

4.1.4. Australia

In the analysis of Australia personality traits relation with emoji usage, Table: 5 it is noticed that there is no significant correlation with "Honesty-Humility and Emotionality". Including "Openness to Experience and Agreeableness" behavior are mostly relevant to use the emoji in online communication. It can be seen that 6 emoji (😀 grinning face, 😅 similing face with open mouth and cold sweat, etc…) explanatory variables significantly affected "Openness to Experience" personality trait behavior. This table describe the some emoji results which have significant relationship with personality traits. Other emoji are omitted based on the correlation results with have weak correlation with personality traits.

4.1.5. Sri Lanka

From Table: 6, it can be seen that 22(out of 65) emoji are significantly association with "Agreeableness" personality trait. This table describe the some emoji results which have significant relationship with personality traits. Other emoji are omitted based on the correlation results with have weak correlation with personality traits.

These emoji were able to explain Sri Lanka peoples' personality traits in online communication. Low conscientiousness describes people more flexible and spontaneous, but also negligent and unreliable, that prefer to they have the "Agreeableness" personality traits. This table describe the some emoji results which have significant relationship with personality traits. Other emoji are omitted based on the correlation results with have weak correlation with personality traits.

5. Conclusion

The objective of this study is to investigate the relationship between HEXACO personality traits and emoji in online communication that emoji could provide information related to personality differences. As per the results are drawn through the country base analysis on the personal traits and emoji that have shown the significant and strong relationship between the HEXACO personality traits. Therefore as an overall analysis, it can conclude that there is a strong relationship between personality traits and emoji when it comes to country base analysis. When taking the countries in isolation, Germany indicating the highest level of correlation and Australia has the lowest level of correlation with personality traits and emoji usage. The usage of emoji across cultures is evident by the fact that variance analyses of difference countries based emoji are statistically significant difference in HEXACO personality traits.

References

[1] Barbieri, F., Kruszewski, G., Ronzano, F., & Saggion, H. (2016). How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 2016 ACM on Multimedia Conference* (pp. 531-535). New York, NY: ACM.

[2] Chen, Z., Lu, X., Shen, S., Ai, W., Liu, X., & Mei, Q. (2017). Through a gender lens: An empirical study of emoji usage over large-scale Android users. Retrieved from http://arxiv.org/abs/1705.05546

[3] Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, NJ: Erlbaum.

[4] Derks, D., Bos, A. E. R., & Grumbkow, J. von. (2007). Emoticons and social interaction on the Internet: The importance of social context. *Computers in Human Behavior*, *23*, 842–849. doi:10.1016/j.chb.2004.11.013

[5] Derks, D., Bos, A. E. R., & von Grumbkow, J. (2008). Emoticons in computer-mediated communication: Social motives and social context. *CyberPsychology & Behavior*, *11*, 99–101.

[6] Glikson, E., Cheshin, A., & van Kleef, G. A. (2017). The dark side of a smiley: Effects of smiling emoticons on virtual first impressions. Social Psychological and Personality Science, 1–12. doi:10.1177/1948550617720269

[7] Jaeger, S. R., Lee, S. M., Kim, K.-O., Chheang, S. L., Jin, D., & Ares, G. (2017). Measurement of product emotions using emoji surveys: Case studies with tasted foods and beverages. Food Quality and Preference, Volume 62, December 2017, Page 46–59.

[8] Prada, M., Rodrigues, D. L., Garrido, M. V., Lopes, D., Cavalheiro, B., & Gaspar, R. (2018). Motives, frequency and attitudes toward emoji and emoticon use. Telematics and Informatics. Telematic and Informatics, volume 35, Issue 7, October 2018, Page 1925-1934

Appendix

Table 2: Emoji based personality scores for Germany

😕	😔	😥	😣	🤔	**H**	.816*	.748*	.759*	.760*	.865**
				😲	**E**	.763*				
	😳	😖	😕	🤤	**X**	-.830*	-.887**	-.870**	-.741*	
	😭	😤	😎	😀	**A**	.796*	-.732*	-.827*	-.862**	
		😫	😪	🤐	**C**	-.737*	-.846**	-.869**		
😱	😟	😋	😆	😁	**O**	-.773*	-.843**	-.785*	-.717*	.726*

Table 3: Emoji based personality scores for Thailand

😕	😕	😝	😐	😋	😉	**H**	-.888**	-.797*	.761*	-.719*	.781*	.714*
				😗	🤐	**E**	-.731*	-.770*				
				😛	😶	**X**	-.943**	.772*				
				😥	😏	**A**	-.755*	-.717*				
😱	😭	😤	😖	😔	😮	**C**	.790*	.835**	-.843**	-.761*	-.731*	-.792*
			☹️	🤑	😁	**O**	-.919**	-.765*	.745*			

Table 4: Emoji based personality scores for Myanmar

H	E	X	A	C	O
😘 -.737*	😀 .832**		😃 .639*	🙄 -.782**	😊 .638*
	😂 .665*		😉 .768**	😯 -.684*	😎 -.654*
	😋 .703*		😍 .813**	😌 .759*	😏 .672*
	🤔 -.699*		😣 -.659*	🤤 -.794**	😒 .711*
	😛 .720*		😢 -.679*	😖 -.817**	😳 .810**
	😔 .834**			😞 -.950**	😡 .681*
	😭 .808**			😟 -.681*	

Table 5: Emoji based personality scores for Australia

						H						
						E						
						**X**	.731*	.676*	.634*	.654*		
						**A**	-.635*	.654*	.814**	.660*	-.633*	
						**C**	-.697*	-.685*	-.705*			
						**O**	.659*	-.673*	-.651*	-.689*	.774**	-.707*

Table 6: Emoji based personality scores for Sri Lanka

H	**E**	**X**	**A**	**C**	**O**
 -.769**	 .682*	 .719*	 -.808**		 .895**
	 .720*	 -.808**	 -.695*		 .817**
			 -.792**		 .644*
			 -.671*		 .843**
			 -.772**		 .771**
			 -.799**		 .836**
			 -.838**		 -.639*
			 -.771**		 .665*
			 -.688*		.851**
			-.821**		.891**
			-.725*		-.723*
			-.740*		-.737*
			-.633*		
			-.671*		
			-.673*		
			-.751*		
			-.738*		
			-.850**		
			-.635*		
			-.730*		
			-.649*		
			-.740*		

University of Computer Studies (Pakokku)
Department of Higher Education
Ministry of Education
Myanmar