Evaluation Of Outlier Detection For Trajectory Data

Nwe Nwe

Computer University (Monywa), Myanmar
Email: nwenwemdy08@gmail.com

Abstract: Outlier of trajectory dataset is different from other in this trajectory dataset. The outlier is involved according to human error, sensors or mechanical faults and system behavior or environment. It becomes challenges in accuracy of clustering, classification and other data mining task. The problem statement is how to detect the outlier and what will be more effective method to detect these outliers. Outlier is detected to increase data quality for all applications. To detect these outliers, similarity measurement is used. In this system, Longest Common Subsequence (LCSS) based measurement and Housdroff Distance (HD) are applied. A comparison result of experimental study on effectiveness of these two methods is described. Experimental observation demonstrates that LCSS Distance produces better results than the other algorithm.

Keywords: clustering, outlier detection, similarity measurement, trajectory data.

1. Introduction

According to the development of the positioning devices, location based services (GPS and sensor) and mobile technology, mass volume of big trajectory data is increase timely. Large amount of data that describe the motion history of moving objects is called trajectory. It is applied in many application domains such as motion detection, human behavior, finding hotspot, traffic pattern analysis, public transportation management, urban planning, location-based service, personalized advertisement and recommendation. It is essential in data mining task. Location can be different dimension, spatial, time, spatial-time, etc. In real application, location is defined by two or more dimensions (longitude, latitude). Smartphone users, surveillance camera generate huge amount of data in rapidly, the big trajectory data become accumulated. To extract the required data, data mining is a key concept especially in clustering, classification and prediction. Clustering is a process of detecting the similar objects and makes a cluster (group). The objects in same cluster have minimum inter-cluster distance. It is very useful to extract meaningful information in data mining. Therefore many research and analysis are applied to trajectory data to cluster the interested group, classify different characteristic of patterns. Clustering requires a distance function to measure the similarity between two trajectories. However, it has been challenges in accuracy of clustering. Accuracy is key concept in all tasks. It is improved by removing of outlier. An outlier is defined as a data object that is different from remaining set of data. Outlier occurs due to many reasons like human error, accuracy of sensor machine and some GPS receivers. Unacceptable ranges of outlier cause the unexpected results. It can reduce accuracy of results. The problem statement is how to detect the outlier and what will be more effective method to detect these outliers. To address this issue, the detecting of outlier is presented by using similarity based on clustering of moving objects in big trajectory data and what will be more effective similarity measurement to detect outliers. Distance based similarity measurement is used to detect the outliers. It is defined as an outlier according to user's defined threshold value. The trajectory data sets will be input and produce required clusters and outlier after processing these data. In this paper, two different similarity measures of Longest Common Subsequence (LCSS) and Housdroff Distance (HD) are presented. Both their performance and their final results are evaluated according to their similarity between the two selected trajectories with their accuracy. This trajectory similarity measurement is important in many applications. Moreover, by using these results, the outliers of clusters have been extracted for frequent pattern mining, data cleaning, and others also. It reaches the higher accuracy and higher data quality. The rest of this paper is organized as follows: section 2 presents the review of the trajectory similarity measure which is used to examine in this work. Section 3 describes the trajectory measurement by using two similarity measurement methods. Section 4 presents the experimental results. Section 5 concludes with future work.

2. Related Work

In this section, the related work is presented in terms of outlier detection in clustering, similarity measure of two different are LCSS and HD methods. Data mining is an important step of process to extract useful information from huge datasets. The main approaches in data mining are clustering, classification and prediction. In clustering, minimum inter-cluster distance exist in same cluster (group). Outlier detection is to point out deviation from others and identified these points as outlier, maximum intra-cluster. Although clustering analysis and outlier detection are different tasks, they are presented together in data mining. To achieve the higher accuracy of clustering, outlier is tested in this paper. A discussion of how distance and similarity functions are described to determine cluster membership which is presented in [1]. A system to detect the outlier and remove the outlier using K-Means and Hierarchical clustering is presented in [2]. The joint clustering with outlier detection problem and also presented clustering with outlier removal algorithm is presented in [3]. The technique for outlier detection which is applicable for very high dimensional datasets has developed in [4]. They found out lower dimensional projections which are locally sparse, and cannot be discovered easily by brute force techniques because of the number of combinations of possibilities. A novel framework, the partititon-and-detect framework, for trajectory outlier detection is proposed in [5]. Incremental local outlier factor (LOF) algorithm for detecting of outlier in data stream is proposed in [6]. A comparative experimental study on the effectiveness of six widely used trajectory similarity measured based on a real taxi trajectory dataset was conducted in [7]. An algorithm for the measurement of the trajectory similarity was implemented for mobile device running on Android OS in [8]. An overview on trajectory outlier detection algorithms from

three perspective, firstly algorithm considering multi-attribute, secondly, suitable distance metric, and thirdly other studies attempt to improve existing algorithm is presented in [9]. Exactly or approximately the smallest Hausdroff distance over all possible rigid motions is computed in [10]. Hausdroff distance is used to compare two binary images is described in [11]. The method compares a 32x32 model bitmap with a 256x256 image bitmap in fraction of a second. Hausdroff distance to express the spatial similarity between two trajectories is applied in [12, 13]. The comparison of outlier detection in big trajectory data using Euclidean distance and Hausdroff distance is presented in [14]. According to these studies, the evaluation of similarity measurement is presented to detect the outlier. To achieve performance of this evaluation, two similarity measurement methods of Longest Common Subsequence and Hausdroff distance are compared. This similarity measurement for outlier detection is useful for clustering, classification and prediction. For example, clustering of data mining task, dataset must be clean to achieve higher data quality. To do so, outlier detection is important to obtain high quality of dataset.

3. TRAJECTORY SIMILARITY MEASUREMENT

3.1 Trajectory distance for outlier

To measure the similarity of trajectory, two similarity methods are applied in this system. They are Hausdroff Distance and Longest Common Subsequence based measurement.

3.1.1 Longest Common subsequence

Longest common subsequence distance to measure similarity is applied in [15]. It is used to obtain the longest common subsequence existing in two trajectory sequences. The longest common subsequence is generally solved recursively, as shown in equation (1).

$$D_L(L_i, L_j) = \begin{cases} 0 & n=m=0 \\ 1+LCSS_{\sigma,\varepsilon}(Head(L_i), Head(L_j)) & |a_i^x - b_j^x| \le \sigma, \text{ and } |a_i^y - b_j^y| \le \varepsilon \\ Max\ (LCSS_{\sigma,\varepsilon}(Head(L_i), Head(L_j)), LCSS_{\sigma,\varepsilon}(Head(L_i), Head(L_j))) & \text{others} \end{cases} \quad (1)$$

where $D_L(L_i, L_j)$ is the longest common subsequence of two trajectory segments L_i and Lj whose length are n and m respectively. The location of moving objects are recorded in 2-dimention. **σ** and **ε** are user defined thresholes of the x-direction and y-direction respectively which are used to control how far when looking for matches. When the abscissa difference and ordinate difference between two trajectories A and B is respectively less than **σ** and **ε**, the pair of trajectory points is considered similar and the value of LCSS is increased by 1. If the numbers of points of trajectory L_i and the numbers of points of trajectory L_j are equal to 0, then $D_L(L_i, L_j)$ is 0.

3.1.2 Hausdroff Distance

Hausdroff Distance measures the similarities by considering how close every point of one trajectory to some points of the other one, and it measure trajectories A and B without unifying the length in [16, 17]. Hausdroff distance is used to find the minimum distance between two trajectories and time-order in data is ignored which is presented by [18].

$$D(A,B) = max\{d(A,B), d(B,A)\} \quad (2)$$

$$\begin{Bmatrix} d(A,B) = max_{a\in A} min_{b\in B} \|a-b\| \\ d(B,A) = max_{b\in B} min_{a\in A} \|b-a\| \end{Bmatrix} \quad (3)$$

3.2 Trajectory dataset

Some dataset has different transformation mode. Therefore two trajectories are applied in this system. They are Geolife trajectory dataset and Taxi dataset. Geolife: This GPS trajectory dataset (version 1.3) was collected in Geolife project by 182 users in a period of over five years (from April 2007 to August 2012). A GPS trajectory of this dataset is represented by a sequence of time-stamped points, each of which contains the information of latitude, longitude and altitude. Transportation modes are walk, bike, bus, car, subway, train, airplane, boat, run and motorcycle. Date and time are displayed in GMT. This trajectory dataset is applied to measure distance of moving objects in [14,19]. The sample of Geolife dataset is presented in table 1.

Table 1 *Sample Geolife Trajectory Dataset*

Data ID	Latitude	Longitude	Altitude	Date and Time
1	39.98379	116.299578	212	2008-10-23, 03:03:00
2	40.013867	116.306473	226	2008-10-23, 23:41:04
3	40.013812	116.306483	156	2008-10-24, 23:44:05

Taxi dataset: The dataset contains 15,054 taxies in Singapore. For each taxi, the GPS information is collected for one entire month with the sampling rates from half a minute to three months. This dataset is applied by [20]. It average distance between two neighbor points is much higher than that in Geolife.

3.3 Data Preprocessing

In the trajectory data analysis, data preprocessing is an essential. In the trajectory dataset, huge amount of data is generated. It contains redundant and noisy information which are recorded and logged every second by GPS or sensor devices. Data collection and cleaning of a set of trajectories are done offline in the data preprocessing. After preprocessing of trajectories, the preprocessed data is used for outlier detection.

4. EXPERIMENTAL RESULTS

We use Java to present this experiment. This system is run on Intel® Core i7, 3.1 GHz with 4GB RAM and 1000 GB hard disk. Experiments are done to determine similarity measure using Longest Common Subsequence distance (LCSS) and Hausdroff distance. Distances between each trajectory and every other trajectory are calculated using these two methods. The minimum distance and maximum distance are also defined. Threshold is set at 0.03 and 0.05

respectively. For each trajectory, cluster of nearest trajectories are formed based on these two methods. If the trajectory has less than threshold value then it is identified as an outlier. The nearest trajectories are clustered in same group. The number of outlier depends on similarity and outlier thereshold value. Moreover, the accuracy and Silhouette index of before and after removing outlier are calculated. The maximum and minimum distance of the two trajectory datasets are shown in table 2. Maximum and minimum values of pairwise distance for all observation are taken as follows.

Threshold value= (maximum distance+ minimum distance)/2 (4)

Table 2 *Results of outlier detection in Longest Common Subsequence Distance and Hausdroff Distance*

No	Dataset	Minimum Distance	Maximum Distance	Threshold Value
1	Geolife Dataset	0.7816	0.1839	0.48275
2	Taxi	0.8732	0.1034	0.4883

The similarity distance using two similarity measurement methods is determined. According the threshold value, the outlier which has smaller distance than thereshold value is defined. Using these outlier values, the accuracy and Silhouette for two trajectory datasets are calculated. The accuracy results for before outlier removal and after outlier removal using Euclidean Distance are presented in table 3, and figure 1 respectively. The accuracy (Acc) of the outlier detection is calculated by using following equation.

$$Acc = (TP + TN)/(TP + TN + FP + FN) \quad (5)$$

where TP= True Positive, TN= True Negative,
FP= False Positive, FN= False Negative.

Table 3 *Accuracy Results for before and after outlier removal using Longest Common Subsequence Distance*

No	Dataset	Accuracy	
		Before	After
1	Geolife Dataset	0.7427	0.7983
2	Taxi	0.7844	0.8783

The Silhouette results for before outlier removal and after outlier removal using Longest Common Subsequence Distance are presented in table 4, and figure 2 respectively.

Table 4 *Silhouette Results for before and after outlier removal using Longest Common Subsequence Distance*

No	Dataset	Silhouette	
		Before	After
1	Geolife Dataset	0.5216	0.6732
2	Taxi	0.5636	0.6851

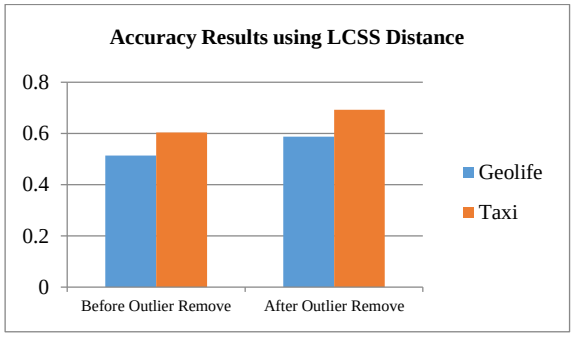

Figure 1 *Accuracy results using Longest Common Subsequence Distance*

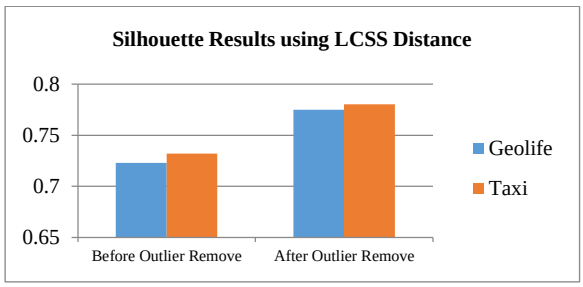

Figure 2 *Silhouette results using Longest Common Subsequence distance*

The accuracy results for before outlier removal and after outlier removal using Hausdroff Distance are presented in table 5, and figure 3 respectively.

Table 5 *Accuracy Results for before and after outlier removal using Hausdroff Distance*

No	Dataset	Accuracy	
		Before	After
1	Geolife Dataset	0.6319	0.6832
2	Taxi	0.7108	0.7633

The Silhouette results for before outlier removal and after outlier removal using Hausdroff Distance are presented in table 6, and figure 4respectively.

Table 6 *Silhouette Results for before and after outlier removal using Hausdroff Distance*

No	Dataset	Silhouette	
		Before	After
1	Geolife Dataset	0.4318	0.4751
2	Taxi	0.5021	0.5418

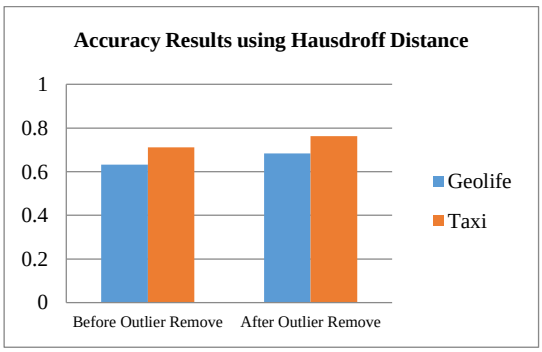

Figure 3 *Accuracy results using Hausdroff Distance*

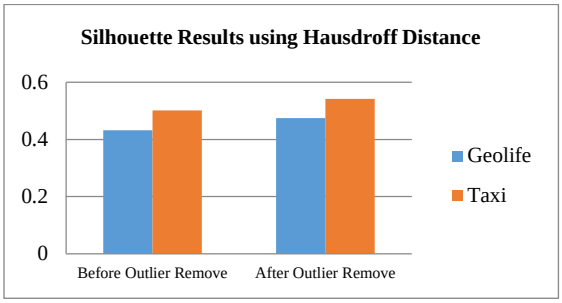

Figure 4 *Silhouette results using Hausdroff Distance*

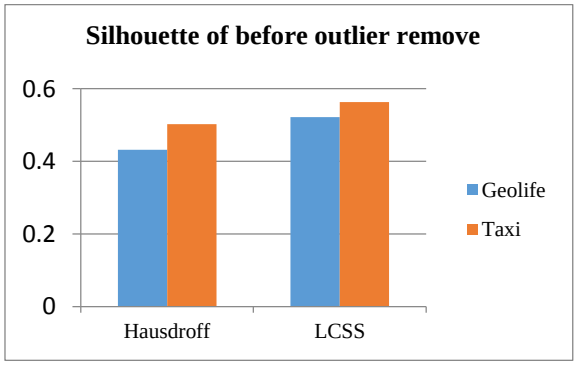

Figure 5 *Silhouette results of before removing outlier*

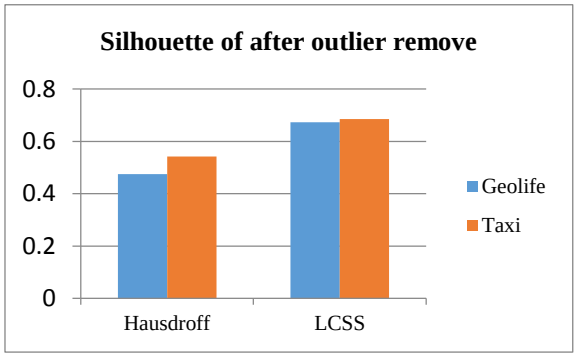

Figure 6 *Silhouette results of after removing outlier*

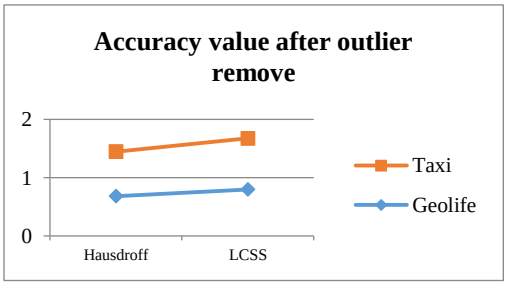

Figure 7 *Accuracy results of after removing outlier*

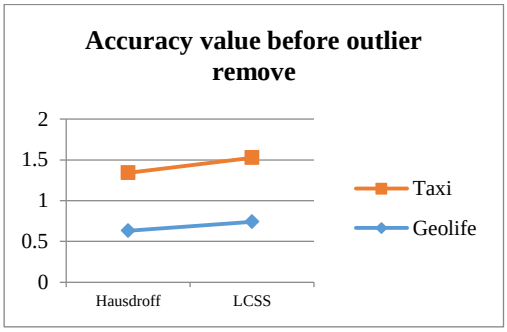

Figure 8 *Accuracy results of before removing outlier*

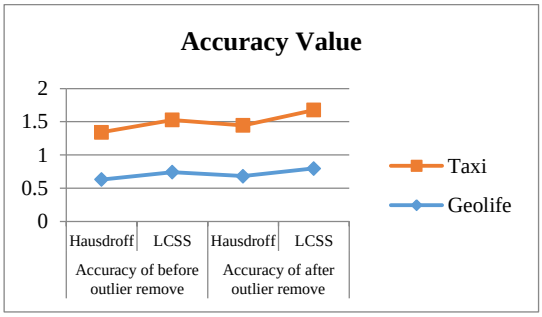

Figure 9 *Accuracy value for two different methods*

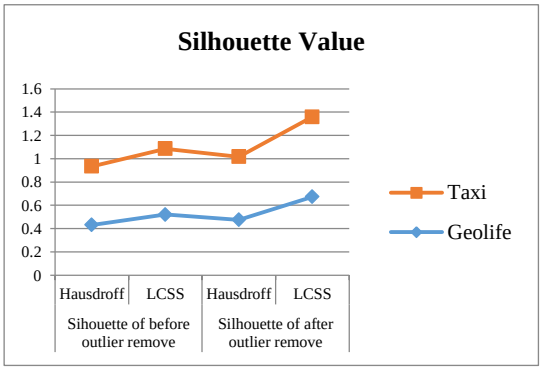

Figure 10 *Silhouette value for two different methods*

The figure 9 and 10 described the accuracy and silhouette value of two different methods. This results show LCSS is good for these experiments. This means LCSS is better for processing of low quality trajectory data to achieve high accuracy. The experimental results of before removing of outlier and after removing of outlier for accuracy and Silhouette are presented in figure 5, 6, 7, and 8 respectively. The comparison results of accuracy value and silhouette value are described in figure 9 and 10 respectively. According to experimental results, after outlier detection, accuracy and silhouette value are increased. The experimental results described that number of outlier depend on defined threshold value. According to this fact, when the threshold value is increase, the number of outlier is also increase in this system. Moreover, other attribute are also considered in next experiments. The experiments are also continued to do with more dataset and more similarity measurement methods. Among them, the most appropriate method for outlier detection can be described.

5. CONCLUSIONS

The two different trajectory similarity methods for outlier detection are compared and the execution time is not considered. According to the experimental results, the number of outlier depends on similarity measurement and outlier threshold. When outlier threshold increase, number of outlier also increase and most of the trajectories in cluster become outliers. Hausdroff distance produces better results than the other algorithm. It is more efficient for minimum distance between two trajectories. The performance and accuracy of cluster is reduced by outlier. According to this fact, the efficient method for outlier detection will be evaluated for data quality in future.

REFERENCES

[1] L. Rokach, "A survey of clustering algorithms" in Data Mining and Knowledge Dicovery Handbook, O. Maimon and L. Rokach, Eds, 2 nd ed. Springer, 2010, pp. 269-298

[2] A. Barai, L. Dey "Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering" in World Journal of Computer Application and Technology, 5(2), 2017, pp. 24-29.

[3] H. Liu, J. Li, Y. Wu and Y. Fu, "Clustering with Outlier Removal" in Proceeding of ACM SIG on Knowledge Discovery and Data Mining (KDD' 18), ACM, 2018

[4] C. Aggarwal, and P. S. Yu, "Outlier detection for high dimensional data' in ACM Sigmod Record, 2001,volume 30, No.2, pp. 37-46.

[5] J. G. Lee, J. HaX. Li, "Trajectory outlier detection: A partition-and-detect framework" in Data Engineering, 2008, ICDE 2008 IEEE 24th International Conference on (pp. 140-149)

[6] D. Pokrajac, A. Lazarevic and L. J. Latecki, "Incremental local outlier detection for data stream' in Computational Intelligence and Data Mining, March, 2007, CIDM 2007, IEEE Symposium on pp. 504-515.

[7] H. Wang, H. Su, K. Zheng, S. Sadiq, X. Zhou, "An Effectiveness Study on Trajectory Similarity Measures", Proceeding of the Tweety-Fourth Australasian Database Conference (ADC 2013), Adlaide, Australia.

[8] N. Larios, C. Mitatakis, V. Kalogeraki, D. Gunopulos, " Evaluation distance measures for trajectories in the mobile setting" proceeding of the 2nd International Workshop on Mining Urban Data, Lille,France, 2015.

[9] F. Meng, G. Yuan, S. Lv, Z. Wang, S. Xia, "An overview on trajectory outlier detection", Springer Science and Business Media, B. v, part of Springer Nature 2018.

[10] P. K. Agarwal, S, Har-Pelet, M. Sharir, Y. Wand, "Hausdrodd distance under translation for points and balls" in Proceeding of the nineteenth annual symposium on Computational geometry, June, 2003, pp. 282-291, ACM.

[11] M. Nanni, D. Pedreschi, "Time-focused clustering of trajectories of moving objects", Journal of Intelligent Information Systems, 27 (3), 2006, pp. 267-289.

[12] J. Lou, Q. Liu, T. Tan, W. Hu, "Semantic Interpretation of Object Activities in a Surveillance System', in Proceeding of 16th International Conference on Pattern Recognition, (ICPR'02) vol3, 2002

[13] I. Junejo, O. Javed, M. Shah, "Multi Feature Path Modeling for Video Surveillance', in Proceeding of 17th International Conference on Pattern Recognition, (ICPR' 04), vol2, pp. 716-719, 2004.

[14] Nwe Nwe, Comparison of outlier detection in big trajectory data", 2018 Joint International Conference on Science, Technology and Innovation, Mandalay by IEEE, 5 October, 2018.

[15] D. Zhang, M. Ding, D. Yang, Y. Liu, J. Fan, H. T. Shen, "Trajectory Simplification: An Experimental Study and Quality Analysis", in Proceeding of the VLDB Endowment, Vol 11, No. 9, 2018.

[16] M. Y. Liu, O. Tuzel, S. Ramalingam, R. Chellappa, "Entropy-rate clustering: Cluster analysis via maximizing a submodular function subject to a matroid constraint' IEEE Transaction on Pattern Analysis and Machine Intelligence, 36(1), 2014, pp. 99-112

[17] J. Chen, R. Wang, L. Liu, J. Song, "Clustering of trajectories based on Hausdorff distance', in Electronics, Communications and Control (ICECC), 2011 International Conference on IEEE, 2011, pp. 1940-1944.

[18] V. Mirge, K. Verma, S. Gupta, "Outlier Detection in Vehicle Trajectories', International Journal of Computer Applications, Vol 171, No. 8, August, 2017.

[19] Khaing Phyo Wai, Nwe Nwe, "Measuring the Distance of Moving Objects from Big Trajectory Data", in 16th International conference in computer and information Science (ICIS) 2017.

[20] J. Bian, D. Tian, Y. Tang, D. Tao, "A survey on trajectory clustering analysis", in Journal of LATEX, February, 2018.

Author Profile

Author 1 received the B.S. and M.S. degrees in Chemistry from Mandalay Arts and Science University in 1992 and 1997, respectively. She achieved Master of Information Science and Ph. D degrees from University of Computer Studies, Yangon in 2000 and 2004 respectively. She is working at Faculty of Information Science of University of Computer Studies, Monywa in Myanmar.