Automated Students' Thai Online Homework Assignment Clustering

Thannicha Thongyoo, King Mongkut's University of Technology North Bangkok (KMUTNB), Thailand
Somkid Saelee, King Mongkut's University of Technology North Bangkok, (KMUTNB), Thailand
Soradech Krootjohn, King Mongkut's University of Technology North Bangkok, (KMUTNB), Thailand

The Asian Conference on Education & International Development 2016
Official Conference Proceedings

Abstract
This paper proposes a model to clustering students' Thai online homework assignments before teachers go further for grading, in other words Automated Students' Thai Homework Assignment Clustering. The proposed model consists of 5 parts: 1) Thai Word segmentation, 2) Stop-word removal, 3) Term Weighting, 4) Document Clustering, and 5) Performance Evaluation**.** The Thai Word segmentation splits sentences into individual tokens. The Stop-word removal defined as a term which is not thought to convey any meaning as a dimension in the vector space. The Term Weighting converts all tokens to vector space by TF-IDF. The Document Clustering is the process that is related to clustering algorithms computing a k-way cluster of a set of other documents. The last element of the prototype is performance evaluation of clustering validation measures in comparison between machines and human beings. The prototype was developed and tested by using Java programming, K-Means Library of Weka 3.7.9, and MySql DBMS. The experiment was conducted with 1,000 undergraduate students who were assigned to complete a particular exercise in the course of Information Technology for Learning. The experimental results showed that the performance of the model could be as effective as the human performance with the 0.92 of Entropy, 0.80 of Purity and 0.80 of F-measure notwithstanding, the model apparently produced similar results significantly to human begins and faster outcome, which can facilitates teachers in terms of clustering into student groups and the students' responses, compared to other students or homework assignment grading.

Keywords: Document Clustering, e-Learning assessment, K-Means

Introduction

Education is great challenge, as a result of technology and its influence in the school. As a result, there are new ways of teaching, new tools for learning, new ways of living and even new ways of thinking. In the 21st Century, the classroom is especially advent of large scale e-Learning. For instance, Massive Open Online Course (MOOC) is an online course aimed at unlimited participation and open access via the web [1]. Teaching activities are different from traditional teaching activities. Students will be taught according to their own abilities and interests. The contents of the lessons which include text, images, audio and video will be delivered to students via a web browser. Students can consult teachers and share their ideas with each other as well as students in regular classes using modern communication tools such as e-mail, web-board, and chat. It is a class for anyone, anywhere and anytime supporting a large number of students. Homework assignment is a task or piece of work assigned to someone as part of a course of study. The students' homework assignment and their submission were provided in the electronic form. Although the e-Learning has a grader, it has a problem in terms of wages and the consistency of graders. The e-Learning classes are likely to increase in term of the number of attending students. One course may have several thousand students enrolled. It takes enormous times in grading students' homework assignment. Teachers do not have time to prepare teaching material or perform a research.

Essays are the most useful tool for assess of learning outcomes. However, it has in terms of the different human assessors, according to Mason [2] reported about 30% of teachers' time is devoted to marking on essays. Researchers were interested in the development and in use of automated assessment tools for essays which have grown exponentially, due to both the increase of the number of students attending universities and to the possibilities provided by e-Learning approaches. The idea of automated essay grading is based on text categorization techniques. Valenti, Neri, & Cucchiarelli [3] reports current tools for automated essay grading such as Project Essay Grade (PEG), Intelligent Essay Assessor (IEA), Educational Testing service I, Electronic Essay Rater (ERater), C-Rater, BETSY, Intelligent Essay Marking System, SEAR, Paperless School free text Marking Engine and Automark. NLP and classification techniques were applied for the current tools for automated essay grading mentioned above. However, no tools for automated Thai essay grading has been done yet.

The advance of technology, results in the immense of amount information in the form of e-document; therefore, document clustering is necessary in grouping information and doing text mining, information retrieval, pattern recognition, and keyword clustering [4], [5], [6], [7]. Document clustering allows these operations to be more efficient in terms of speed.

According to the problems mentioned above, the researcher is interested in developing a model to analyze students' Homework assignment clustering using K-Means technique**.** It is hoped that this model can help teachers in grouping students' Homework assignment more effectively with shorter time.

Literature Review

Document Clustering systems are used more and more often in text mining, especially to analyze texts and to extract knowledge they contain [8][9]. For example, Business: Market research companies use clustering a lot. With the clusters defined, the marketing companies can try to develop new products or think about testing products for certain clusters in the results. The Internet: Social media network analysis uses clustering to determine communities of users. Regarding to Computing: With the rise of the "Internet of Things". Clustering can be used to group the results of the sensors. Course work in the education sector, especially with the advent of large scale learning online, can be clustered into student groups and results. Clustering is used often in digital imaging. When large groups of images need to be segmented, it's usually a cluster algorithm that works on the set and defines the clusters. Algorithms can be trained to recognize faces, specific objects, or borders. Law Enforcement: crimes are logged with all the aspects of the felony listed. Police departments are running clustering and other machine learning algorithms to predict when and where future crimes will happen.

The Clustering are several algorithms in clustering data which are difficult to define the best one because each algorithm has its own strengths and weaknesses. Several researchers compared the effectiveness of algorithms such as Abbas [10] who studied and compared the differences of 4 data clustering algorithms, i.e. K-Means, HCA, SOM, and EM. All these algorithms are compared according to the following factors: size of dataset, number of clusters, type of dataset and type of software used. He conclude that the performance of K-Means and EM are better than HCA and SOM, the quality of K-Means and EM become very good for huge dataset. Amine, Elberrichi, and Simonet [11] conducted a research on evaluation of text clustering methods using WordNet. The results obtained show that the SOM-based clustering method using the cosine distance provides the best results. Karnjana [12] proposed a new method of data clustering called K-Inverse harmonic means (KIHM) by using inverse radial basis function to calculate the interval instead of Euclidian distance measurement. She conclude that the data clustering with accuracy from the highest to the lowest as follows IHM, KHM, KM, and when considering the performance of each method has descending order of the KHM, KIHM, KM. Kantiga [13] compared the effectiveness of data clustering and concluded that SOM together with Fuzzy C-Means resulted in better effectiveness when each clustering data overlapped. Kwale [14] conducted a study on K-Means and family including K-Means, K-Medians, Bisecting K-Means and K-Medoids (PAM, CLARA, CLARANS) to find strengths and weaknesses. He can conclude about K-Means and family is easy to use effectively with few weaknesses. Besides, he suggested that it should be used together with other document clustering algorithm to select the strength of certain manner. There are several researchers who are interested in the effectiveness of using clustering algorithm and have applied in several works.

Kanokrat and Nattanon [15] applied data clustering technique to investigate the research on the analysis of specialists' opinions: Inference analysis and data clustering for Delphi Technique researchers. They analyzed the inferences by clustering words and phrases with the same structure level using bi-setting K-Means clustering technique. They conclude that their developed model can help reduce analyzing time and errors from bias. Albayrak and Amasyali [16] have developed medical diagnosis

system using clustering technique in grouping data of thyroid patients. Chureerat, Jetsada, and Sataporn [17] have employed clustering technique in categorizing Thai news. Oranuch [18] conducted a research by grouping Thai handicraft customers by Kohonen's Self Organizing Maps (SOM) together with K-Means and found that they were appropriate in dealing with a high number of data while Hierarchical Cluster (HC) together with K-Means are appropriate for dealing with a low number of data. Sasithorn [19] using SOM with Fuzzy C-Means for grouped Technology Transfer and Agricultural Service centers in local area of Thailand.

As mentioned in the literature review above, clustering technique has been widely used in grouping documents and other applications. However, there has been no application of document clustering in analyzing students' Thai homework assignment. The researcher has an applying K-Means algorithm for clustering students' Thai online homework assignment.

Research Methodology

The researcher has a concept in Automated Students' Online Homework Assignment Clustering to help teachers to grouping students' Thai online homework assignment effectively with shorter time. The conceptual framework of the research is shown in Figure 1. Principles of Automated Students' Thai Online Homework Assignment Clustering can be described as follows. When the teacher assigned the homework assignment via e-Learning system, after students finished their homework assignment, they sent it on line in the system. The Automated Students' Thai Online Homework Assignment Clustering will group answers of students according to the similarity of the documents before submitting them to the teacher grading.

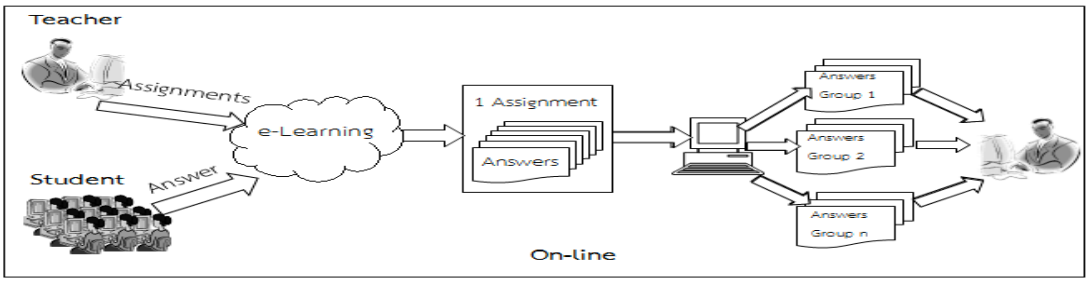

Figure 1: A conceptual framework of Automated Students' Thai Online Homework assignment Clustering.

1) Research Method

The proposed model of Automated Students' Thai Online Homework Assignment Clustering consists of 5 parts: 1) Thai Word segmentation, 2) Stop-word removal, 3) Term Weighting, 4) Document Clustering, and 5) Performance Evaluation**.** Each part is processed sequentially as shown in Figure 2. The Thai Word Segmentation that splits sentences into individual tokens. The Stop-word removal defined as a term, which is not thought to convey any meaning as a dimension in the vector space. The Term Weighting converts all tokens to vector space by TFIDF. The Document Clustering is the main focus of this research, the process that is related to clustering algorithms computing a k-way cluster of a set of other documents. The last element of the prototype is performance evaluation of clustering validation measures in comparison between machines and human beings.

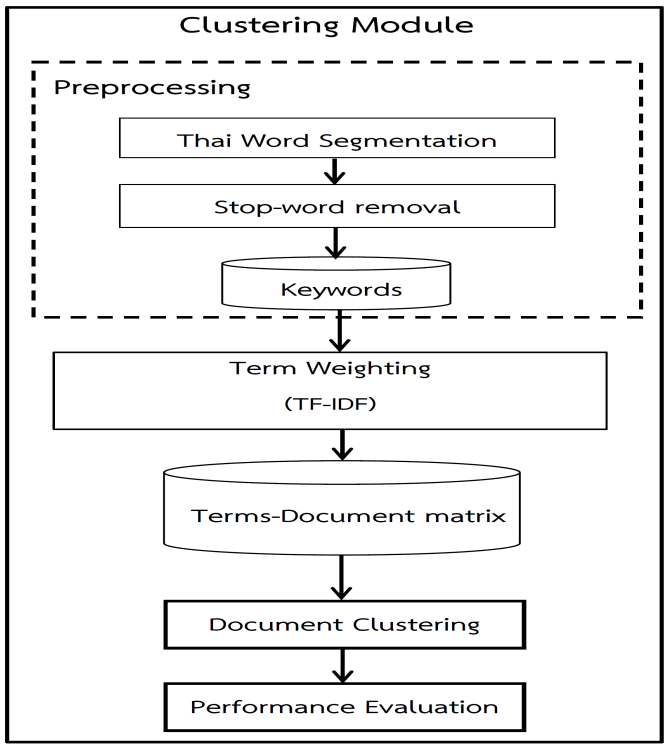

Figure 2: The proposed model of Automated Students' Thai Online Homework Assignment Clustering.

The detail of proposed model of each process is explained and illustrated in Figure 3.

Document	Thai Word Segmentation
Doc1	ค้นคว้า|ความรู้|เพิ่มเติม|จาก|เนื้อหา|การเรียน|ใน|ห้องเรียน
Doc2	ค้นคว้า|ข้อมูล|ใน|เรื่อง|ที่|เรียน|และ|สิ่ง|ที่|อยาก|ได้|ความรู้|เพิ่มเติม
Doc3	ช่วย|ใน|การ|ค้นคว้า|ข้อมูล|ได้|สะดวก|และ|มี|ประสิทธิภาพ|มาก|ยิ่งขึ้น
Doc4	ใช้|ติดต่อ|สื่อสาร|กับ|เพื่อน|กับ|อาจารย์|ได้|สะดวก|มาก|ขึ้น
…..	……

Document	Stop-word removal
Doc1	ค้นคว้า|ความรู้|~~เพิ่มเติม|จาก~~|เนื้อหา|การเรียน|~~ใน|~~ห้องเรียน
Doc2	ค้นคว้า|ข้อมูล|~~ใน|เรื่อง|ที่|~~เรียน|~~และ|สิ่ง|ที่|อยาก|ได้~~|ความรู้|~~เพิ่มเติม~~
Doc3	~~ช่วย|ใน|การ~~|ค้นคว้า|ข้อมูล|~~ได้|สะดวก|และ|มี~~|ประสิทธิภาพ|~~มาก|ยิ่งขึ้น~~
Doc4	~~ใช้|ติดต่อ~~|สื่อสาร|~~กับ~~|เพื่อน|~~กับ~~|อาจารย์|~~ได้|สะดวก|มาก|ขึ้น~~
…..	……

Term Weighting

Key word	tf_{ij}				df_i	N/df_i	IDF_i	$W_{ij}=tf_{ij}*idf_i$			
	N1	N2	N3	N4				N1	N2	N3	N4
ค้นคว้า	1	1	1	1	4	1	0	0	0	0	0
ข้อมูล	3	0	1	0	2	2	0.3	0.9	0	0.3	0
ความรู้	0	0	1	1	2	2	0.3	0	0	0.3	0.3
อาจารย์	0	1	0	2	2	2	0.3	0	0.3	0	0.6
สื่อสาร	1	1	2	1	4	1	0	0	0	0	0

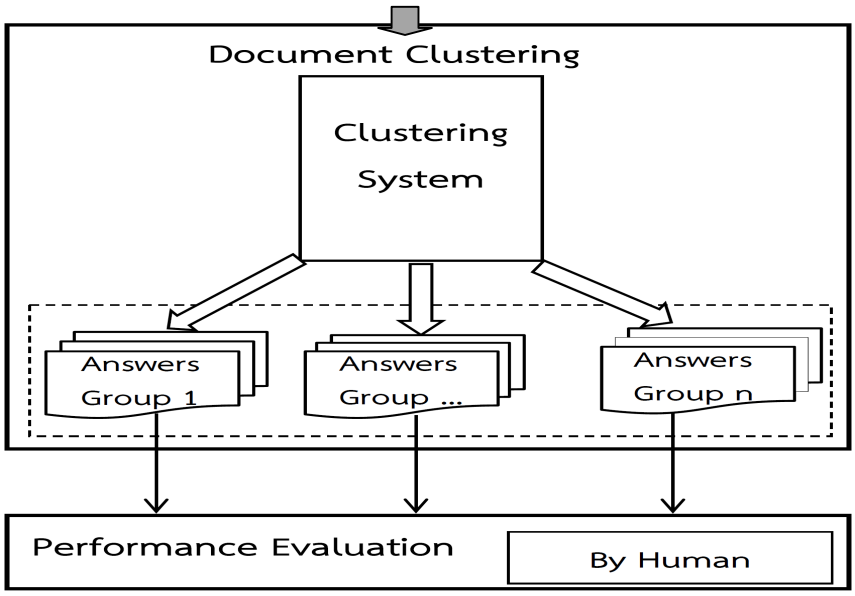

Figure 3: The detail of proposed model of Analysis of Automated Students' Thai Online Homework Assignment Clustering.

2) Population and Sample

The research was conducted with 1000 undergraduate students, registering for the subject of Information Technology for Learning at Thepsatri Rajabhat University. Each answer document contains less than 250 words. The label or evaluation clustering had been done by 23 teachers.

3) Evaluation Instruments

The evaluation was done by using the Entropy, Purity, Recall, Precision and F-measure [20] are calculated as follows.

-The entropy of a cluster

$$E_i = \sum_i -(P_{ij}) \log_2 (P_{ij})$$

- The overall entropy

$$E = \sum_{i=1}^{r} \frac{n_i}{n} * E_i$$

- The purity of a cluster

$$purity_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\}$$

- The overall purity

$$purity = \sum_{i=1}^{r} \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^{r} \max_{j=1}^{k} \{n_{ij}\}$$

-The F-measure of a cluster

$$F_{ij} = \frac{2 * \mathrm{Re}call(i,i) * \mathrm{Pr}ecision(i,j)}{\mathrm{Re}call(i,i) + \mathrm{Pr}ecision(i,j)}$$

$$\mathrm{Recall}(i, j) = \frac{n_{ij}}{n_j}$$

$$\mathrm{Precision}(i, j) = \frac{n_{ij}}{n_i}$$

-The overall F-measure

$$F = \sum_{i=1}^{r} \frac{n_j}{n} * \max \{F(i,j)\}$$

4) Data preparation

The subjects in this study were asked to complete the exercise in the course of Information Technology for Learning with 1000 answers. The preprocessing had been done on Thai Word Segmentation and Stop-word removal before Thai Word Segmentation split the sentences into individual tokens. The Stop-word removal was defined as a term, which was not conveyed any meaning as a dimension in the vector space using stop-word dictionary.

5) Research tools and algorithm for clustering

The prototype of Analysis of Students' Thai Online Homework Assignment Clustering was developed and tested by using Java programming, K-Means Library of Weka 3.7.9, and MySql DBMS. The answers were clustered into groups of similar documents using K-Means algorithm. The main advantages are that K-means clustering is easy to use and understand, and works faster and more efficiently with smaller documents, and uses less memory O*(k)* and with less time complexity O(*knl*): whereas, *n* is the number of patterns, *k* is the number of clusters, and *l* is the number of iterations taken by the algorithm to converge [21].

Experimental Results

The experiments used K-Means algorithm for clustering by configuring k = 2, 3, 4, and 5. The tables below are present the performance of document grouping using K-Means algorithm. The research findings are described in details as follows:

Table 1: Measures of cluster Validity (2 classes).

Label	By Human		Total	Entropy	Purity	Recall	Precision	F-measure
By Machine	Cluster #0	Cluster #1						
Cluster #0	**296**	40	336	0.53	0.88	0.93	0.88	0.90
Cluster #1	23	**641**	664	0.22	0.97	0.94	0.97	0.95
Total	319	681	1,000	0.32	0.94	0.93	0.92	0.94

Table 2: Measures of Cluster Validity (3 classes).

Label	By Human			Total	Entropy	Purity	Recall	Precision	F-
By Machine	Cluster #0	Cluster #1	Cluster #2						measure
Cluster #0	295	12	35	342	0.69	0.86	0.86	0.86	0.86
Cluster #1	21	360	40	421	0.73	0.86	0.94	0.86	0.90
Cluster #2	27	9	201	237	0.74	0.85	0.73	0.85	0.78
Total	343	381	276	1,000	0.72	0.86	0.84	0.86	0.85

Table 3: Measures of Cluster Validity (4 classes).

Label	By Human				Total	Entropy	Purity	Recall	Precision	F-
By Machine	Cluster #0	Cluster #1	Cluster #2	Cluster #3						measur
Cluster #0	**361**	28	38	24	451	1.03	0.80	0.89	0.80	0.84
Cluster #1	8	**78**	11	8	105	1.23	0.74	0.48	0.74	0.58
Cluster #2	26	17	**105**	45	193	1.67	0.54	0.64	0.54	0.59
Cluster #3	12	39	11	**189**	251	1.13	0.75	0.71	0.75	0.73
Total	407	162	165	266	1,000	1.20	0.73	0.68	0.71	0.73

Table 4: Measures of Cluster Validity (5 classes).

Label	By Human					Total	Entropy	Purity	Recall	Precision	F-
By Machine	Cluster #0	Cluster #1	Cluster #2	Cluster #3	Cluster #4						measure
Cluster	**146**	9	27	28	51	261	1.78	0.56	0.72	0.56	0.63
Cluster	12	**125**	22	14	25	198	1.66	0.63	0.80	0.63	0.70
Cluster	12	6	**78**	7	17	120	1.59	0.65	0.51	0.65	0.57
Cluster	14	6	19	**98**	8	145	1.51	0.68	0.65	0.68	0.66
Cluster	19	11	6	4	**236**	276	0.85	0.86	0.70	0.86	0.77
Total	203	157	152	151	337	1,000	1.44	0.68	0.68	0.67	0.68

Table5: The average efficiency rating of Automated Students' Thai Homework Assignment Clustering.

No. of Cluster	Entropy	Purity	Recall	Precision	F-measure
2 Cluster	0.32	0.94	0.93	0.92	0.94
3 Cluster	0.72	0.86	0.84	0.86	0.85
4 Cluster	1.20	0.73	0.68	0.71	0.73
5 Cluster	1.44	0.68	0.68	0.67	0.68
Average	**0.92**	**0.80**	**0.78**	**0.79**	**0.80**

The results showed that the developed model can find the similar or the same answers accurately with less time than human work.

Conclusion and Future Work

The purpose of this research was to group students' Thai Online Homework Assignment using document clustering. Students participated in the experiment consisted of 1000 bachelor degree students registered in Information Technology for Learning. The assignments were administered to the students and the students' answers were clustered. The experimental results showed that the performance of the model could be as effective as the human performance with the 0.92 of Entropy, 0.80

of Purity and 0.80 of F-measure; notwithstanding, the model apparently produced similar results significantly to human begins and faster outcome.

According to the research finding, this research showed that document clustering technique can grouping the documents correctly with similar results done manually. However, document clustering took less time compared to human. It can be grouping homework assignment before grading. It clearly clustering can perform reduce the time in grading the homework assignment. The teachers have more time to develop other innovative works such as research or new teaching medias and help them reducing employment of graders in institutions.

There are still several ways in which our work can be enhanced. Based on our results, we plan to use classification techniques to develop automated Thai homework assignment grading system.

References

Faculty of Science, Mahidol University. (2014). "MOOCs". Stang Mongkolsuk library and Information division newsletter. [cited Aug 20, 2014] . Available from : http://stang.sc.mahidol.ac.th/newsletter/pdf/apr57-1.pdf.

Mason, O., and Grove-Stephensen, I. (2002). Automated free text marking with paperless school. In Proceedings of the 21st ACM/SIGIR(SIGIR-98), 90-96. ACM.

Valenti, S., Neri, F., and Cucchiarelli, A. (2003). An overview of current research on automated essay grading. Journal of Information Technology Education: Research, *2*(1), 319-330.

Han, J., Kamber, M., and Pei, J. (2006). Data mining: concepts and techniques. Morgan kaufmann.

Maimon, O. Z., and Rokach, L. (Eds.). (2010). Data mining and knowledge discovery handbook (Vol. 2). New York: Springer.

Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

Supachai Tangwongsan .(2010). Information Storage and Retrieval System. 2nd ed. Bangkok: Pitakkanpim.

Bell, J. (2015). Machine Learning: Hands-On for Developers and Technical Professionals. John Wiley & Sons, Inc.

Michael J.A. Berry and Gordon S. Linoff. (2004). Data mining techniques : for marketing, sales, and customer relationship management. 2nd ed. Wiley Publishing.

Abbas, O. A. (2008). Comparisons Between Data Clustering Algorithms. Int. Arab J. Inf. Technol., 5(3), 320-325.

Amine, A., Elberrichi, Z., and Simonet, M. (2010). Evaluation of text clustering methods using wordnet. Int. Arab J. Inf. Technol., 7(4), 349-357.

Karnjana Siripaisam. (2011). K-Inverse Harmonic Means Clustering Algorithm. KKU Res J (GS), 11(2). 21-30.

Kantiga Promm. (2013). Comparison of Two-stage Clustering Algorithms . National Graduate Research Conference(NGRC29) . 294-301.

Kwale, F. M. (2013). A Critical Review of K Means Text Clustering Algorithms. Internaltional Jouranl of Advanced Research in Computer Science, 4(9).

Kanokrat Jirasatchanukul and Nattanon Hongwarittorn. (2011). Analyze System of Expert's Opinions using Latent Semantic Analysis and Text Clustering Technique. Information Technology Journal Vol. 8, No. 1, January – June 2012.

Albayrak, S. and Amasyali, F. (2003). Fuzzy c-means clustering on Medical Diagnostic Systems. Proceedings of the 12th International Turkish Symposium on Artificial Intelligence and Neural Networks.

Chureerat Jaraskulchai, Jetsada Gantasena, and Sataporn Kewsuwannasuk. (2001). Thai Text Document Clustering. Proc. of 5th National Computer Science and Engineering Conference (NCSEC 2001), 7-9 Nov., Chiangmai University, Thailand.

Oranuch Chaimuen. (2005). Comparison of Clustering of Thai Handcraft Customers by Using Two Stage Clustering: SOM and K-Means Algorithm and Hierarchical Clustering and K-Means Algorithm. Master's thesis, Kasetsart University, Faculty of Science and Technology, Department of Computer Science.

Sasithorn Mongkonsipattana. (2005). Clustering of Center for Technology Transfer and Services of Agricultural district in Thailand by Using Two Stage Clustering: SOM and Fuzzy C-Means Algorithm. Master's thesis, Kasetsart University, Faculty of Science and Technology, Department of Computer Science.

Mohammed J. Zaki, Wagner Meira, Jr. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

Contact email: thongyoo99@hotmail.com