SNRU Journal of Science and Technology

Journal home page: snrujst.snru.ac.th

Improvement on Automated Thai Assignment Scoring by Using a Thesaurus

Thannicha Thongyoo*, Somkid Saelee, Soradech Krootjohn

Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok

***Corresponding Author**: thongyoo99@hotmail.com

Received: 15 October 2017; **Revised:** 30 November 2017; **Accepted:** 4 December 2017; **Available online:** 1 January 2018

Abstract

Essays are a great useful tool to assess students' learning outcomes. Automated essay scoring represents a practical solution to the time-consuming activity of manual scoring of students' essays. However, one problem of the efficiency automated essay scoring is the synonym. The objective of this research was to compare the results of an experimental study of automated Thai assignment scoring with a thesaurus to solve the synonym problem and without a thesaurus. The experiment was conducted with 1,000 undergraduate students who were assigned to complete a particular on-line exercise in the course of Information Technology for Learning at the Thepsatri Rajabhat University, Lopburi, Thailand. The proposed model has 2 types of categories, Thai Assignment Scoring without Thesaurus (TASWOT) and Thai Assignment Scoring with Thesaurus (TASWT). Both types were designed using the principles of text mining including Document Clustering and Document Classification. The accuracy rate of classification was used to evaluate efficiency. The experimental results showed that the performance of the automated Thai assignment scoring with a thesaurus equaled 83.77 percent of average accuracy, while the automated Thai assignment scoring without a thesaurus had 74.92 percent of average accuracy. The results of the experiment indicated that automated Thai assignment scoring with a thesaurus had higher efficiency than automated Thai assignment scoring without a thesaurus. Because, that the use of Thesaurus to solve the synonym problems before clustering has resulted in some documents shifted into other clusters when comparing with clustering without a Thesaurus. As a result, machine scoring has been changed and that also leads to the effectiveness of machine scoring getting better.

Keywords: automated Thai assignment scoring; thesaurus; document clustering; document classification; machine learning

1. Introduction

Essays are a great useful tool to assess students' learning outcomes. However, scoring students' essays takes an enormous time, labor-intensive and expensive activity for educational institutions, according to Mason's reported [1] about 30% of teachers' time is devoted to scoring essays. Teachers do not have much time to prepare teaching materials or conduct their research, so time is very valuable. 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 aims at unlimited participation and open access via the web [2]. It takes enormous time in grading students' assignments, although the MOOC has graders, it has a problem in terms of wages and the consistency of graders. A practical solution to many problems associated with manual scoring is to have an automated system for essay evaluation. Researchers are interested in the development of automated assessment tools for essays, due to both the increase of the number of students attending universities and the possibilities provided by e-Learning approaches. The idea of automated essay scoring is based on text categorization techniques. Valenti, et al. [3] reported current tools for automated essay scoring 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 have been applied to the current tools for automated essay scoring mentioned above, but it has not been using clustering techniques and has not been done in automated Thai assignment scoring. Clustering techniques [4 – 6] are used more and more often in Document Retrieval, Data Mining, Object and Character Recognition, Image Segmentation, especially to analyze texts and to extract knowledge they contain. Clustering techniques enhance the effectiveness of these operations in terms of speed. Abbas [7] studied and compared the differences of data clustering algorithms, and concluded that the performance of K-Means and EM are better than HCA and SOM, The quality of K-Means and EM becomes very good for huge datasets. Kwale [8] conducted a study on K-Means and its family including K-Means, K-Medians, Bisecting K-Means, and K-Medoids (PAM, CLARA, CLARANS) to find strengths and weaknesses. He concluded that K-Means and family were easy to used and effective with few weaknesses. Several researchers are interested in the effectiveness of using the K-Means clustering algorithm and have applied it in several works. For example, Jirasatchanukul and Hongwarittorn [9] applied data clustering techniques to investigate 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 the bi-setting K-Means clustering technique. They concluded that their developed model could help reduce analyzing time and errors from bias. Chaimuen [10] conducted a study by grouping Thai handicraft customers by using Kohonen's Self Organizing Maps (SOM) and K-Means, the result showed that they were appropriate in dealing with a high number of data while Hierarchical Cluster (HC) together with K-Means were appropriate for dealing with a low number of data. Junwipat [11] applied SVM for data categorization with a Thai hand-written consonant recognition program. The results obtained show that the SVM can achieve a high accuracy rate, more than 86%, on real datasets. Chirawichitchai, Sanguansat, and Meesad [12] conducted a research on Thai document categorization with SVM by adjusting kernel function parameters. They concluded from their experimental results that reducing the features with an information gain method and process with SVM (Linear kernel) and SVM (Polynomial kernel Degree = 3) yielded a very high classification with the F-Measurement equal to 95.10%. For this reason, Thongyoo, Saelee, and Krootjohn [13] have applied clustering techniques and classification techniques to Automated Students' Thai Online Homework Assignments Scoring. Based on this research, researchers found that the difference between human scoring and machine scoring is the humans often interpret the contents while scoring. In addition, they have some related matters in terms of mind, emotions, and other environments as the main factors affecting the scoring process, all of which cannot be seen in the scoring process operated by the machine. However, one problem of the efficiency automated essay scoring is the synonym. A synonym [14] is a word or phrase that means nearly the same as another word or phrase in the same language. There are many synonyms in the Thai language. For example, the words "ค้น", "ค้นหา", "สืบค้น", "หา" are synonym. One solution to solve problems associated with synonyms is to have a thesaurus. The thesaurus [15] is a collection of selected vocabulary (preferred terms or descriptors) with links among Synonymous, Equivalent, Broader, Narrower and other Related Terms. The words are derived from: 1) The specialists in each field of study, or 2) All the words were categorized by the researcher. After that, grouping is performed to create the relationships. One of the most outstanding features of the thesaurus is that all words are related. It consisted of 5 parts [16]: 1) Broader Term: BT, 2) Narrower Term: NT, 3) Related Term: RT, 4) Use For: UF, and 5) USE. For example [17], คอมพิวเตอร์ BT เครื่องจักรประมวลผล,

คอมพิวเตอร์ NT การประมวลผลข้อมูลอิเล็กทรอกนิกส์, คอมพิวเตอร์ RT ซอฟต์แวร์, คอมพิวเตอร์ UF คณิตกร, *คณิตกร USE คอมพิวเตอร์*

According to the problems mentioned above, this paper focused on developing a model for automated Thai assignment scoring using clustering techniques with classification techniques and comparing its efficiency of the experimental study of automated Thai assignment scoring without a thesaurus and with a thesaurus.

The basic outline of this paper is as follows: Section II provides a proposed model, tools for development and implementation. The Results and Discussion are explained in section III. Section IV describes the Conclusion. Section V consists of Suggestions.

2. Materials and methods

Proposed model

The overview of the research was applied to identify the students' responses. To illustrate the processes of the system, teacher assigned the assignments via e-Learning system and students submitted their assignments online in the system. The system grouped the answers from students according to the similarity of the documents into cluster. Each cluster eventually was classified by the SVM algorithm to score and returned the scores to the teacher (See Fig. 1).

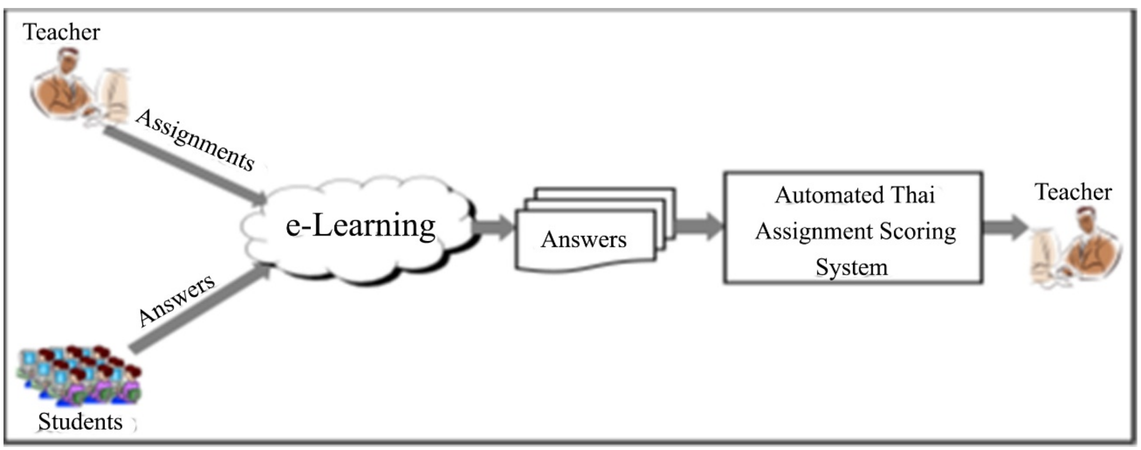

Fig. 1 Overview of the research.

The objective of this research was to compare the results of an experimental study of Automated Thai assignment scoring without a thesaurus and with a thesaurus to solve the synonym problem. The proposed model has 2 types of categories, Thai Assignment Scoring without Thesaurus (TASWOT) and Thai Assignment Scoring with Thesaurus (TASWT). Both types were designed using the principles of text mining including Document Clustering and Document Classification (See Fig. 2).

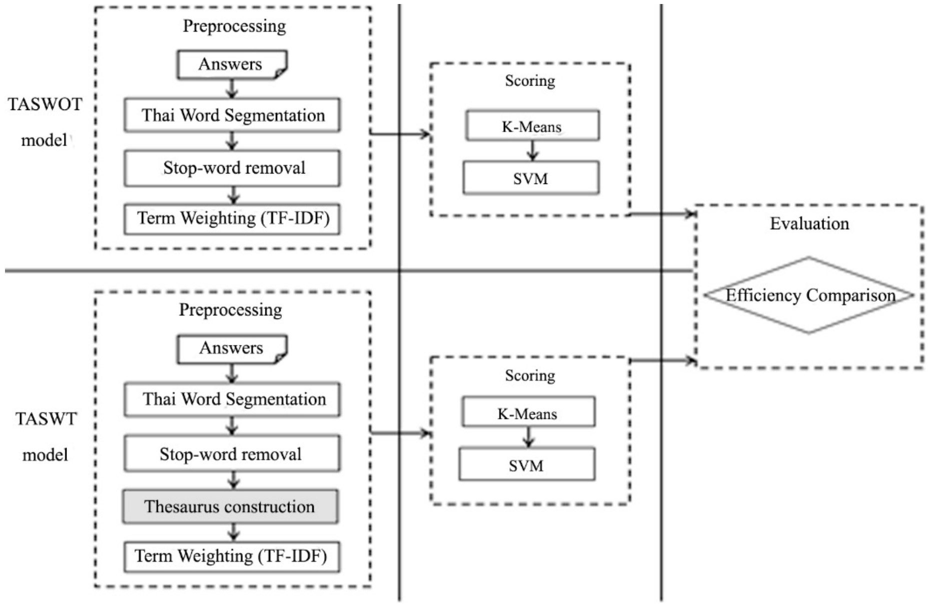

Fig. 2 Proposed Model.

The difference detail of the two types is the TASWOT model does not use thesaurus (See Fig. 3), while the TASWT model uses a thesaurus to solve synonym problems (See Fig. 4).

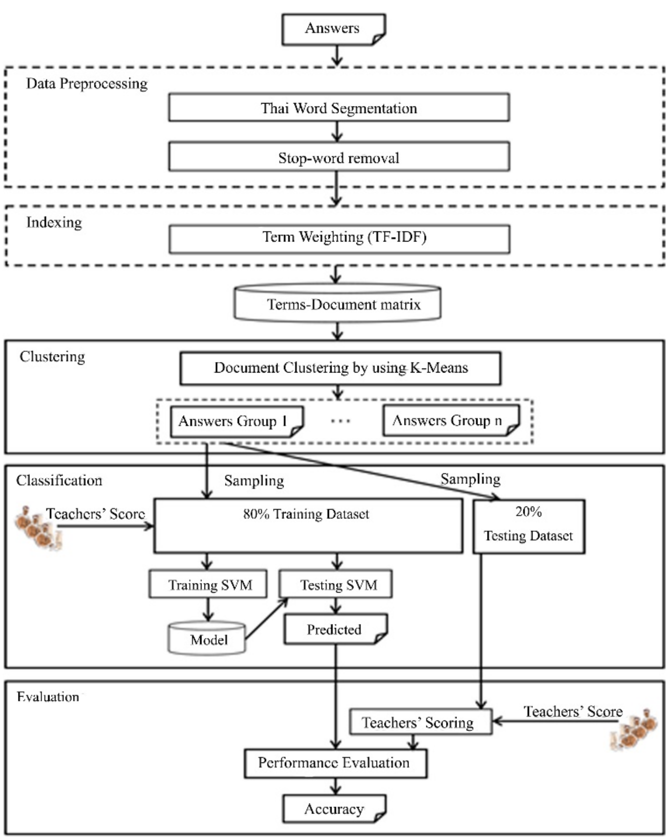

Fig. 3 TASWOT architecture.

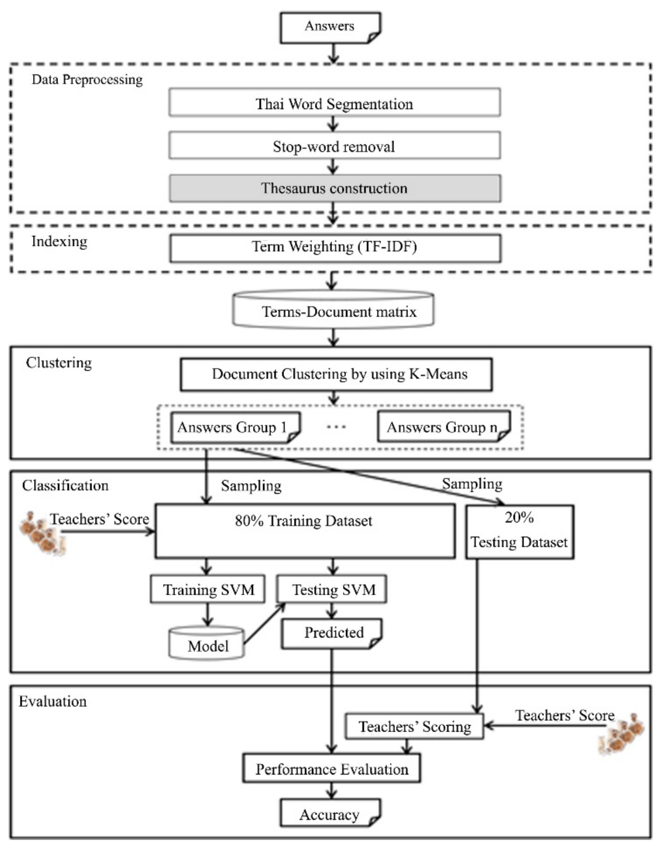

Fig. 4 TASWT architecture.

Tools for development

The used platform is the Windows 7 operating system. The development environment is NetBeans IDE 8.0.2 and Java programming. The software of K-Means is Weka 3.7.9 [18]. The software of SVM is LIBSVM [19].

Implementation

The proposed model consists of 5 parts: 1) Data Preprocessing, 2) Term Weighting, 3) Document Clustering, 4) Document Classification, and 5) Performance Evaluation. The sequential process of the model is described as follows:

Data Preprocessing

The data included 1,000 answers by sampled the answers of students according to stratified sampling [20] from the undergraduate students who registered for the course of Information Technology for Learning at Thepsatri Rajabhat University, Lopburi, Thailand. When the teacher assigned the assignments via e-Learning system and the students submitted their assignments online in the system. Each answer contained less than 250 words. The answers were graded by three teaching experts on the subject of Information Technology for Learning. The popular vote (Mode) of the human grades was used for training and performance evaluation. Data pre-processing is the feature extraction that is used to recognize and classify significant vocabulary items in natural language texts. The steps of the feature extraction are removed empty answers, Thai Word Segmentation, Stop-word removal, and the construction of Thesaurus. Thai Word Segmentation split the sentences into individual tokens using LexTo software, developed by National Electronics and Computer Technology Center (NECTEC) [21]. LexTo software is Thai word segmentation using the longest matching technique. The Stop-word removal was defined as a term, which did not convey any meaning as a dimension in the vector space using a stop-word dictionary. Thesaurus is used to solve synonym problems for the TASWT model. The flow of work in Thesaurus construction for this research is shown in Fig. 5.

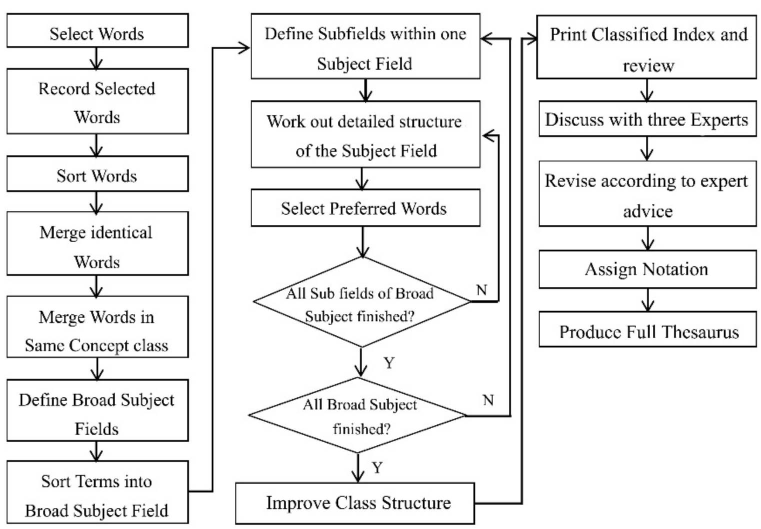

Fig. 5 Flow of work in Thesaurus construction.

Indexing

The Indexing is the feature selection by converting all words into vector space for the use of clustering, training SVM, and testing the SVM. The feature selection that were conducted by using TF-IDF consists of 4 steps: 1) term frequency (TF), which computes frequency of words, 2) document frequency (DF) which computes of document frequency of term in the collection, 3) inverse document frequency (IDF), which inverses the proportion of the word over the entire document collection, and 4) term weighting (TF-IDF), which computes the weight of words. The formula of TF-IDF [16] is shown in equation (1) – equation (4) as follows;

$$TF = tf_{ij} = frequency\ of\ term\ i\ in\ document\ j \quad (1)$$

$$DF = df_i = document\ frequency\ of\ term\ i\ in\ the\ collection \quad (2)$$

$$IDF = idf_{ij} = log\ (N/df_i) \quad (3)$$

$$TF\text{-}IDF = tf_{ij} * idf_{ij} \quad (4)$$

where tf_{ij} is term frequency of term **i** in document j, df_i is document frequency of term i in the collection, idf_{ij} is inverse document frequency of term i in the collection and N is the number of document in the collection.

Document Clustering

Consequently, the answers were clustered into groups of similar documents using the K-Means algorithm [4 – 6, 22 – 24]. K-Means algorithm is presented as in the following process:

1) Input Number (K) of Cluster.
2) Select K points as the initial centroids.
3) Repeat
    3.1) Form K clusters by assigning all points to the closest centroid.
    3.2) Recompute the centroid of each cluster until the centroids stop changing.

The main advantages are that K-means clustering [19] is easy to use and understand, works faster and more efficiently with smaller documents, uses less memory O(k) and 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. In this process, it begins with a teacher providing a value of k, which is determined by the scores of assignments. For instance, if the full marks are five, k will also be set as five, then system will select the best seed to solve the problem of disadvantage of the K-Means algorithm, if randomly selecting the inappropriate seed, it will result in poor performance.

Document Classification

Support Vector Machine (SVM) [5, 22 – 24, 26] is supervises the learning method so as to separates the binary classes with a maximized margin criterion called the Optimal Separating Hyperplane. For a training set (x_1, y_1) ... (x_n, y_n) with labels y_i in +1 or -1 and n be the input dimension. The classification can be calculated in equation (5).

$$y_i = \begin{cases} +1;\ if\ (w * x) + b > 0 \\ -1;\ if\ (w * x) + b < 0 \end{cases} \quad (5)$$

where w is weight and b is bias.

Kernel functions help the SVM in finding the optimal solution. The most frequently used such functions are the linear kernel, polynomial kernel, sigmoid kernel and radial basis kernel function (RBF). In real-world problems often require the discrimination for more than two categories. In practice, the multi-class classification problems (k > 2) perform based on binary SVM. Two representative ensemble schemes are one-against-all (1-v-a) and one-against-one (1-v-1) [27] approaches.

One-against-all (1-v-a) or one-versus-rest (1-v-r), this method is used to consider the one class against all other classes by setting the answer y as follows:

$$y_i = \begin{cases} +1, & x_i \in C_k \\ -1, & x_i \notin C_k \end{cases} \tag{6}$$

where C is a class that is considered each time by 1 <= k <= K and will be considered as K times for all classes.

One-against-one (1-v-1) or one-versus-one or Pairwise SVM with configuring an answer y of binary SVM by consider only class C_j and class C_k as follows:

$$y_i = \begin{cases} +1, & x_i \in C_j \\ -1, & x_i \in C_k \end{cases} \tag{7}$$

where C_j is a class that is a comparision with other class by $j \neq k$ and comparision with all class by combination K(K-1)/2 times.

The process of the Document Classification was conducted to score students' assignments using SVM which is currently gaining a great interest for problem solving referring to data classification and pattern recognition [13]. The system built the training file and test file with the 5-fold cross validation of 80 percent of the training dataset and 20 percent of the testing dataset. The training dataset of each cluster was assigned a score ranging from 0-5 points by 3 teaching experts who taught the subject of Information Technology for Learning of Thepsatri Rajabhat University. After that, set value of SVM type and kernel type to build the training model and test dataset for predicted result by used default value of parameters were adjusted to the parameter of SVM as follows: SVM type = C-SVC, cost = 1, weight = 1, probability estimates = 0, kernel function = Radius Basic (RBF), gamma = (1/ number of features) and cache Size = 100. The reason why parameter was used as default value is because this experiment does not aim to use parameter in order to get the best efficient model, but focuses on the comparison of the effectiveness of TASWOR model and TASWT model.

Performance Evaluation

The last element of the model is the performance evaluation which the accuracy of classification will be determined by calculating both correctness and incorrectness in comparison between machine and human performances. The evaluation will be conducted by computing the accuracy of the results [28].

3. Results and Discussion

The experiments used the K-Means algorithm for clustering by configuring k = 5, used SVM algorithm for classification by configuring SVM type = C-SVC, cost = 1, weight = 1, probability estimates = 0, kernel function = Radius Basic (RBF), gamma = (1/ number of features) and cache Size = 100. Moreover, the 5-fold cross validation was used to test the model. The research findings are described in Table I demonstrates the comparison of results of efficiency from TASWOT and TASWT in the course of Information Technology for Learning.

Table 1 The efficiency of TASWOT and TASWT.

Cluster #	TASWOT				TASWT			
	No. of data	Correctly	Incorrect	Accuracy (%)	No. of data	Correctly	Incorrect	Accuracy (%)
0	605	436	169	72.07	497	405	92	81.49
1	48	35	13	72.92	89	73	16	82.02
2	60	42	18	70.00	57	49	8	85.96
3	39	33	6	84.62	92	78	14	84.78
4	48	36	12	75.00	65	55	10	84.62
			Average	74.92		**Average**		83.77

The TASWOT model had 74.92 percent of average accuracy, while the TASWT model had 83.77 percent of average accuracy. When considering the experimental results of each document, it was indicated that the use of Thesaurus to solve the synonym problems before clustering has resulted in some documents shifted into other clusters when comparing with clustering without a Thesaurus. As a result, machine scoring has been changed and that also leads to the effectiveness of machine scoring getting better but in some cases after the problem of synonyms was solved by Thesaurus, the document was then shifted to another group, and the scoring changed from correct to wrong because the human grades was used for training model and performance evaluation that were performed by cluster.

4. Conclusion

The purpose of this research is to compare the results of an experimental study of Automated Thai assignment scoring without a thesaurus and with a thesaurus. The participants taking part in the experiment consisted of 1,000 answers which were sampled by stratified sampling from the undergraduate students, registered in the course of Information Technology for Learning at Thepsatri Rajabhat University, Lopburi, Thailand. The assignments were assigned to the students and the students' answers were clustered and scored, respectively. The experimental results showed that the TASWT had higher efficiency than TASWOT. According to the research findings, the developed model (TASWT) and the use of the thesaurus to automate assignment scoring could reduce the number of similarity in terms of Thai vocabulary with more efficiency for automated assignment scoring.

5. Suggestions

According to the research findings, a thesaurus can solve synonym problems with higher efficiency than Thai assignment scoring without a thesaurus. For future research, based on the results, an adaptive learning model will be developed with more efficiency to accommodate the large scale of e-Learning in the 21st Century, such as MOOC.

6. Acknowledgement

We would like to express our special thanks to Dr. Choochart Haruechaiyasak for his advice on the experiment. Our thanks go to NRCT for granting this research fund and the lecturers at Thepsatri Rajabhat University for their scoring in the experiment. We also appreciate the students' collaboration at Thepsatri Rajabhat University in the experiment.

7. References

[1] O. Mason, I. Grove-Stephensen, Automated free text marking with paperless school, Proceedings of the 6th CAA Conference, Loughborough University. 09 July 2002, 211 – 219.

[2] MOOCs, http://stang.sc.mahidol.ac.th/newsletter/pdf/apr57-1.pdf, 20 August 2015.

[3] S. Valenti, F. Neri, and A. Cucchiarelli, An overview of current research on automated essay grading, JITE:Research. 2(1) (2003) 319 – 330.
[4] S. Tangwongsan, Information Storage and Retrieval System, third ed., Pitakkanpim, Bangkok, 2015.
[5] J. Bell, Machine learning: Hands-on for developers and technical professionals, John Wiley & Sons Inc., Indianapolis, 2015.
[6] A.K. Jain, M.N. Murty, P.J. Flynn, Data Clustering: A Review, CSUR. 31(3) (1999) 265 – 323.
[7] O.A. Abbas, Comparisons Between Data Clustering Algorithms, Int. Arab J. Inf. Technol. 5(3) (2008) 320 – 325.
[8] F. M. Kwale, A Critical Review of K Means Text Clustering Algorithms, IJARCS. 4(9) (2013) 27 – 35.
[9] K. Jirasatjanukul, N. Hongwarittron, Analyze System of Expert's Opinions using Latent Semantic Analysis and Text Clustering Technique, IT Journal, 8(1) (2012) 1 – 12.
[10] O. Chaimuen, Comparison of Clustering of Thai Handcraft Customers by Using Two Stage Clustering: SOM and K-Means Algorithm and Hierarchical Clustering and K-Means Algorithm, M.Sc. Computer Science, Kasetsart University, Bangkok, 2005.
[11] S. Junwipat, Data Classification Problem Solving using Support Vector Machine Approach and Its Application, JIT. 7(1) (2011) 50 – 55.
[12] N. Chirawichitchai, P. Sanguansat, P. Meesad, Automatic Thai Document Categorization with Support Vector Machines, The 6th National Conference on Computing and Information Technology, Bangkok. 3 – 5 June 2010, 87 – 92.
[13] T. Thongyoo, S. Saelee, S. Krootjohn, Automated Thai Online Assignment Scoring, 2016 Fifth ICT International Student Project Conference (ICT-ISPC), Mahidol University. 27 – 28 May 2016, 33 – 36.
[14] Synonym, https://en.wikipedia.org/wiki/Synonym, 20 April 2016.
[15] D. Soegel, Indexing Languages and Thesauri: Construction and Maintenance, Los Angeles: Melville Publishing Co., California, 1974.
[16] U. Thonghuapai, Information Retrieval, Ramkhamhaeng University Press, Bangkok, 2008.
[17] Taxonomy & Thesaurus, http://technology.in.th/thesaurus/index.php?tema=9114&/aaaaaaaaaa-computers, 1 May 2016.
[18] The WEKA Data Mining Software, http://www.cs.waikato.ac.nz/m, 5 August 2013.
[19] LIBSVM: A library for support vector machines, http://www.csie.ntu.edu.tw/cjlin/libsvm, 25 October 2014.
[20] M. Teianthong, Research Methodology in Computer Education, Danex inter-corporation Co.,Ltd, Bangkok, 2012.
[21] LexTo software, http://www.sansarn.com/lexto, 7 May 2014.
[22] O. Maimon, L. Rokach, Data mining and knowledge discovery handbook, second ed., Springer Scienece+Business Media, New York, 2010.
[23] J. Han, M. Kamber, J. Pei, Data mining: concepts and techniques, third ed., Elsevier Inc., Waltham Massachusetts, 2011.
[24] M. Kantardzic, Data mining: concepts, models, methods, and algorithms, second ed., John Wiley & Sons Inc., New Jersey, 2011.
[25] J. MAO, A.K. JAIN, A self-organizing network for hyperellipsoidal clustering (HEC), IEEE Trans. Neural Netw. 7 (1996) 16 – 29.
[26] C.C. Aggarwal, Data classification: algorithms and applications, CRC Press, New York, 2014.
[27] P. Sanguansat, Machine Learning, second ed., Textbook production center of KMUTNB, Bangkok, 2015.
[28] N. Chirawichitchai, Automatic Thai Document Classification Model, JIT. 9(1) (2013) 141 – 149.