Classification of Thai Independent Study in Statistics Using Data Mining Techniques

Phimphaka Taninpong[1*], Nattira Muangmala[2]

Received: 18 March 2013 Accepted: 15 May 2013

Abstract

In this paper, the empirical study of the classification of Thai independent study in statistics is discussed. Our purpose is to classify the undergraduate independent study researches into three groups: sample survey, statistical analysis, and operational research and related field. Several classification techniques, such as support vector machine, Naïve Bayesian, Decision Tree, k-Nearest Neighbor and RBF network, are used in this paper. We also employed the feature selection techniques in order to find the best subset of features that help improve the accuracy of the classification model. The experimental results show that the RBF network algorithm gives a best accuracy when the Chi-square is employed as the feature selection method.

Keywords: Document Classification, Independent Study, Data Mining, Text Mining

Introduction

Nowadays, huges volume of documents are stored in the database system and can be retrieved via the internet such as theses and dissertations, electronic books, news, emails, etc. Since the documents are continuously increasing, categorization of those documents is required in order to improve the efficiency of document retrieval. Text categorization helps to automatically assign the category to a document that its category is unknown. In the data mining perspective, text categorization is also called text classification which employs the statistical learning method, machine learning technique to build the classification model. Consequently, the model will be used to assign the category to a new document. However, text classification has several challenging problems. First, text is sparse data which has a high dimensionality feature space. As the features are words in the document, the feature space may contain several hundreds to thousands of terms. Second, the contents of the document are overlapped and it is difficult for us to determine the separation line between the categories. In our department of statistics, the independent studies in statistics of the undergraduate students were stored in the database system and can be retrieved via the web application. Since the year 2010, we categorized independent study of the undergraduate students into three groups: Sample Survey, Statistical Analysis, and Operational Research and related field. And, they are already classified into three groups using manual classification. However, the independent study researches proposed during the year 2004-2009 have not been classified into three groups. Thus, the objectives of this study are two fold: (1) build the automatic classification model for classifying the Thai independent study reports and (2) classify the Thai undergraduate independent study in statistics during the year 2004-2009 into three groups. In addition, the best classification model will be used to classify automatically the independent study in statistics in the future in order to avoid the manual classification.

The organization of this paper is as follows. The related work on the Thai document classification is given in Section 2 and our research methodology is described in Section 3. Experiments and new results are obtained in Section 4 and Conclusion and future work are discussed in Section 5.

[1,2] *Department of Statistics Faculty of Science, Chiang Mai University 239, Muang District, Chiang Mai 50200, Thailand*

Email : p.taninpong@gmail.com

** Coressponding author: E-mail: p.taninpong@gmail.com*

Related Work

A document is unstructured data in which data is the text. Thus, the text classification techniques is used to classify such text document. Text classification techniques which are widely used including Support Vector Machine[1], Naïve Bayesian[2], etc. However, most of them are applied to the English document classification. For Thai document classification,[3,4,5] presented the comparative study of the impact of feature selection method and the data mining algorithm to an automatic classification of Thai document. Their results show that the support vector machine give the highest accuracy as reported in[2]. They employed the feature selection method in order to reduce the processing time while preserve the accuracy of the model. And, the results show that the information gain help improve the efficiency of the mdel. Our work is similar to[3,4,5] in such the way that we present the empirical study of the classification of Thai independent study by using various feature selection method and data mining algorithm.

Methodology

Figure 1 shows the framework of the classification of Thai independent study classification which requires three processes: (1) preprocessing helps prepare the data before performing the classification, this process consists of tokenization, stop word removing, feature scoring, feature selection, (2) training process constructs the classification model using the preprocessed data, (3) testing process uses the classification model to classify the test data. The detail of each step is elaborately described in the following subsection.

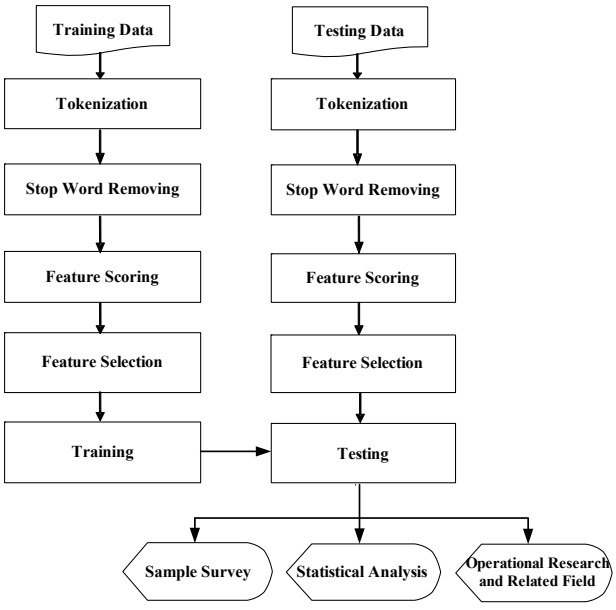

Figure 1 Framework of the classification of Thai Independent Study

A. Preprocessing

Each independent study document consists of title, abstract, contents, etc. In this work, we consider only the title of the independent study. For training data, each title is preprocessed before it is used for learning. The testing data is also preprocessed before it is classified. Since, titles are unstructured text data, the features are referred to words in the title. The preprocessing process consists of four steps: tokenization, stop word removing, feature scoring and feature selection. The detail of each step is described below.

1. Tokenization

Tokenization is the process that breaks the stream of characters into words or tokens. The token delimiters could be character spaces, tabs and newlines, which are not counted as tokens. For Thai text, we employed the SWATH program which is developed by[6] in order to separate the titles into several words. Figure 2 shows an example of Thai text segmentation.

Title	Tokenization
การสำรวจความคิดเห็นของหัวหน้าครัวเรือน การศึกษาตัวแบบการแจกแจงความน่าจะเป็น การศึกษาการถดถอยสำหรับข้อมูลวงกลม	การ|สำรวจ|ความ|คิดเห็น|ของ|หัวหน้า|ครัวเรือน| การ|ศึกษา|ตัว|แบบ|การ|แจกแจง|ความ|น่า|จะ|เป็น| การ|ศึกษา|การ|ถดถอย|สำหรับ|ข้อมูล|วงกลม|

Figure 2 Example of Thai text segmentation

2. Stop Word Removing

Stop words removing eliminates non significant words such as "ณ" (at), "กู" (me), "ฉัน" (I), etc. In this work, we used a list of Thai stop words which is proposed by[7]. Figure 3 shows an example of stop words removing.

Tokenization	Stop word are removed
การ|สำรวจ|ความ|คิดเห็น|ของ|หัวหน้า|ครัวเรือน| การ|ศึกษา|ตัว|แบบ|การ|แจกแจง|ความ|น่า|จะ|เป็น| การ|ศึกษา|การ|ถดถอย|สำหรับ|ข้อมูล|วงกลม|	สำรวจ|คิดเห็น|หัวหน้า|ครัวเรือน| ศึกษา|แจกแจง| ศึกษา|ถดถอย|ข้อมูล|วงกลม|

Figure 3 Example of Stop Word Removing

3. Feature Scoring

Statistically, titles can be represented as a vector space model which is a vector of weighted word frequencies such as term frequency, term frequency-inverse document frequency, or a binary value showing the existence of a word. Term frequency (*tf*) is the number of each word's occurrence in a document. Term frequency-inverse document frequency (*tf-idf*) can be used to compute weighting of words. The *tf-idf* weight assigned to word *j* is the term frequency (*tf*) proportioned by a scale factor according to the word *j*'s importance. The scale factor is called the inverse document frequency, read[8] for more detail.

In this work, we use the binary value to simplify the manipulation of categorical data and eliminate the need for data normalization. However, the feature space can be represented in the form of word by document matrix and it is depicted in Figure 4.

Topic of IS	**Word**				
	สำรวจ	**คิดเห็น**	**หัวหน้า**	**ครัวเรือน**	**...**
1	y	y	n	n	
2	n	n	y	y	
3	y	n	n	n	
4	y	y	y	n	
5	n	n	n	y	
...					

Figure 4 Word by Document Matrix

B. Feature Selection Methods

As more number of documents, more number of words are extracted and the feature space could contain more than several hundreds to thousands words. A high dimensionality feature space has a lot impact to the processing time as well as the accuracy of the classification model. The feature selection method is required to select only the effective words for classification. The feature selection methods which are widely used including information gain (IG), chi-square (CHI), gain ratio, etc.

Information gain measures the number of bits of information obtained for category by knowing the presence or absence of word in a title. The information gain of word *w*, IG(*w*), is defined in (1) below.

$$\mathrm{IG(w)} = -\sum_{j=1}^{k} \mathrm{P(c_j)}\log \mathrm{P(c_j)} + \mathrm{P(w)}\sum_{j=1}^{k} \mathrm{P(c_j \mid w)}\log \mathrm{P(c_j \mid w)} + \mathrm{P(\overline{w})}\sum_{j=1}^{k} \mathrm{P(c_j \mid \overline{w})}\log \mathrm{P(c_j \mid \overline{w})} \quad (1)$$

where $P(c_j)$ is the probability that class c_j was observed in the dataset, P(*w*) is the probability that word *w* occurs in the dataset whereas $P(\overline{w})$ is the probability that word *w* does not occur in the dataset. $P(c_j|w)$ is the probability that class c_j will contain word *w*.

Gain ratio is an extension of information gain which selects words that have maximized the ratio of its gain divided by its entropy[9]. The gain ratio of word *w* is defined in (2):

$$Gain(w) = \frac{H(class) - H(class|w)}{H(w)} \quad (2)$$

where H is the entropy.

Chi-square statistics measures the lack of independence between word *w* and class c_j. The detail of using chi-square statistics to compute the goodness of word for classification is described in[10].

C. Classification Techniques

The classification technique is the supervised learning technique that learns from the dataset which each instance has already been classified. In this work, we employed Neural Network, Support Vector Machine, Naïve Bayesian, Decision tree, k-Nearest Neighbor algorithm. The detail of each algorithm is described as follows.

Neural network is commonly used in supervised learning. A simple neural network structure which consists of the input, hidden and output layers. The input and hidden layers can have multiple nodes, but there will be only a single output. The basic function is to sum up the values of its inputs, and transform them with a function to produce the output.

Multilayer neural nets use the output of single perceptrons as inputs to the subsequent perceptrons. In other words, the outputs of each perceptron are the inputs of the next layer, and all layers between the first layer and the last layer are called the hidden layer. This allows the system to learn more complex features. In this work, we employed the RBF network in Weka[16] which is the neural network that uses the radial basis function as the activation function.

Support Vector Machine (SVM) is a robust machine learning methodology which shows high performance on text classification[1]. The basic concept is to find two hyperplanes that separate two classes of data in data space while maximizing the margin between them.

The SVM can be constructed as a linear or non-linear model. Given that the training dataset X contains n labeled sample vectors $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where each x_i is a feature vector of the document i and each y_i is the class label of the document i. The linear SVM uses a weight vector w and a bias term b to classify a new example x, by creating a predicted class label f(x) as given in (3) below.

$$f(x) = sign(\langle w, x \rangle + b) \tag{3}$$

For the non-separable case, the training errors are allowed so that the linear SVM finds the vector w by minimizing the objective function over all n training samples as shown in (4).

$$T(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=i}^{n} \xi_i \tag{4}$$

under the constraints that

$$\forall i = \{1...n\} : y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \xi_i \geq 0$$

In this work, we employed the Platt's SMO algorithm[11,12,13] in Weka with default parameter for building the support vector machine classification model.

Naïve Bayesian employed in this work is the Naïve Bayesian with nominal attributes[14] and we used NaiveBayes in Weka with default parameter. Equation (5) shows how to predict the class of a testing document and (6) shows how to calculate the probability to indicate whether a document is in class c.

$$c(\vec{x}) = \arg\max_{c \in C} P(C = c)\prod_{i=1}^{n} P(X_i = x_i | C = c) \tag{5}$$

where n is the number of words in the dataset, N_c is the number of documents in a class c, and N is the total number of documents in the training data.

The probability of the word w_i would be in the class c can be defined as:

$$P(X_i = x_i | C = c) = \frac{N_{ic}}{N_c}, \tag{6}$$

where C is the class, N_{ic} is the number of documents in class c that word x_i occurs, N_c is the total number of documents in class c.

Decision Tree techniques find the classification rules based on the tree structure. The decision tree consists of internal node, or so-called non-leaf node, and terminal node, or so-called leaf node. Each internal node denotes a test on a word, each branch represents an outcome of the test[8]. The root node is an internal node which is the best splitting word. And, each leaf node has a class label. The algorithm is repeated to find the best splitting word until a given subset contains documents of only one class. Finally, the classification rules are induced from the final decision tree. In this work, we used J48 in Weka as it implements C4.5[15] which is the well known decision tree algorithm.

The k-Nearest Neighbor technique finds the k closest documents to the testing document by measuring the distance between documents. There are many ways to measure the distance for determining the similarity between documents. In this work, the Euclidean distance is used and k is set to 3 since it gave the best accuracy in

our experiments. If the voting scheme is used, the testing document will be assigned the class label which is the majority class of *k*-Nearest Neighbor.

Experimental Results

In this work, three experiments are conducted using two datasets in order to build the classification model and test the model. The detail of the experimental setup including datasets, and the evaluation metrics are described below.

A. Experimental Setup

1. DataSet: two datasets are described as follows.

a) First dataset: This dataset is used to train the classification model and test the model using 10 folds cross validation. It contains the title of the undergraduate independent study in statistics during the year 2010-2012.

b) Second dataset: This dataset contains the titles of the undergraduate independent study in statistics during the year 2004-2009. This dataset did not used to build the classification model because the research category has not been assigned to each independent study but we aim to group this dataset into three groups using the model constructed from the first dataset.

Table 1 shows the statistics of the number of titles in each group and the total number of words of the independent study dataset for each year.

Table 1 number of the titles, words and research catego

Year	Number of Titles				Number of Words
	Sample Survey	Statistical Analysis	OR and Related field	Total	
2004	N/A	N/A	N/A	19	123
2005	N/A	N/A	N/A	26	139
2006	N/A	N/A	N/A	39	205
2007	N/A	N/A	N/A	45	252
2008	N/A	N/A	N/A	56	291
2009	N/A	N/A	N/A	58	338
2010	32	20	18	70	660
2011	31	31	11	73	478
2012	23	19	14	56	478

2. Evaluation Metrics: Several evaluation measures are used to compare the classification performance of different learning methods. The basic measures are accuracy, true positive, and false positives.

a) Accuracy: the percentage of all titles which are correctly classified.

b) True Positive of class j: the percentage of correctly classified titles for class j, where j=1,2,3.

c) False Positive of class j: the percentage of titles which are not in the class j and incorrectly classified as class j, where j=1,2,3.

B. Experiment I: Investigating the accuracy of the learning method.

This experiment aims to investigate the classification results of the learning method using the training dataset. In this work, we also compare the classification results using the various feature selection methods. The goodness of the words for the classification are mea sured by the feature selection method, and the score are ranked in the descending order. In this work, we selected only top *k* features for classification and the number of selected features (*k*) is varied from 10 to 40 percent of the total features with the increment by 10. Table II shows that the feature selection method has a slightly impact to the accuracy of the learning method since there was a small variation of the accuracy between each feature selection methods. In this work, we assessed the different of the accuracy when using the various feature selection methods by the analysis of the variance (ANOVA). The result shows that there was no significant difference in the accuracy of the various feature selection methods at significance level of 0.05. Table 2 also shows that the RBF learning algorithm gave the highest accuracy when the Chi-square was used as the feature selection method. In addition, the experimental result shows that the RBF algorithm with 30 percent of total features gave the best results.

Table 2 accuracy (%) of the learning methods

Feature Reduction/ Number of features	SVM	NB	RBF	DT	kNN
Without using feature Reduction	90.45	86.93	75.88	81.41	79.40
Chi-Square					
85 (10%)	90.45	86.93	89.95	85.93	80.90
170 (20%)	87.44	90.45	91.46	83.42	80.90
255 (30%)	87.94	89.95	**92.96**	83.92	82.41
340 (40%)	88.44	88.44	91.96	82.41	78.89
Average	88.57	88.94	91.58	83.92	80.78
Gain Ratio					
85 (10%)	84.42	84.42	86.93	83.92	78.39
170 (20%)	86.93	84.42	89.45	85.93	78.89
255 (30%)	88.44	87.44	90.45	82.41	84.42
340 (40%)	88.44	88.44	87.44	86.43	84.42
Average	87.06	86.18	88.57	84.67	81.53
Information Gain					
85 (10%)	88.44	86.43	88.44	85.93	80.40
170 (20%)	90.95	90.95	91.46	86.43	81.91
255 (30%)	88.44	89.45	90.45	82.41	84.42
340 (40%)	87.94	90.45	89.95	82.41	79.90
Average	88.94	89.32	90.08	84.30	81.66

Table 3 classification table

Predict	True		
	Sample Survey	STAT Analysis	Operation Research and related field
Sample Survey	84	6	1
STAT Analysis	1	62	3
Operation Research and related field	1	2	39

Table 3 shows the classification table and the result shows that most of the sample survey and operational research topics are correctly classified. Table 4 shows that the overall accuracy of the model is 92.96 percent and most of the sample survey researches are correctly classified since the TP rate is 97.67 percent whereas the statistical analysis researches are misclassified more than 10 percent.

Table 4 testing results

Group	TP Rate	FP rate
Sample Survey	97.67	6.20
STAT Analysis	88.57	3.10
Operation Research and related field	90.69	1.92
Overall	**92.96**	**11.10**

C. Experiment II: Investigating the impact of the feature space representation.

Since features are lost during tokenization and stop word removing, for example, the word "ความน่าจะเป็น" which should be in the extracted feature list is lost. Thus, we combined single word into word bigram which is a pair of consecutive words. The objective of this experiment is to compare the accuracy of the learning methods between the single word representation and word bigram representation. Table 5 shows the accuracy of the learning methods using word bigram representation and the result shows that the RBF learning algorithm also gived the highest accuracy when the chi-square was used as the feature selection method. In addition, the experimental result shows that the RBF algorithm with 20 percent of total features gives the best result. Table 6 shows that the accuracy of using single word as feature representation is higher than that of using word bigram representation.

Table 5 accuracy (%) of the learning methods

Feature Reduction/ Number of features	SVM	NB	RBF	DT	3NN
Without using feature Reduction	87.44	83.92	74.87	81.41	74.87
Chi-Square					
240 (10%)	87.44	87.44	88.94	82.91	75.88
480 (20%)	87.94	90.95	90.95	83.42	81.91
720 (30%)	86.93	88.94	88.44	83.42	74.37
960 (40%)	87.44	88.44	83.42	81.91	73.37
Average	87.44	88.94	87.94	82.92	76.38
Gain Ratio					
240 (10%)	78.89	77.39	79.40	78.89	70.35
480 (20%)	82.41	81.91	80.90	82.91	70.35
720 (30%)	88.94	83.42	87.44	83.42	73.37
960 (40%)	85.93	75.38	77.89	83.42	67.34
Average	84.04	79.53	81.41	82.16	70.35
Information Gain					
240 (10%)	89.45	87.44	90.45	82.41	77.89
480 (20%)	87.44	90.45	90.95	83.41	83.92
720 (30%)	89.45	88.44	86.93	82.41	79.40
960 (40%)	87.94	87.94	81.41	82.41	76.88
Average	88.57	88.57	87.44	82.66	79.52

Table 6 comparison of accuracy (%) between feature representation using single word and word bigram

Feature Representation	SVM	NB	RBF	DT	3NN
Single Word	87.94	89.95	92.96	83.92	82.41
Word Bigram	87.94	90.95	90.95	83.42	81.91

D. Experiment III: Assign class to the Thai independent Study of Undergraduate in Statistics during the year 2004-2009.

This experiment aims to automatic assign the class label to the second dataset. In this experiment, we employed the RBF algorithm with 30 percent of total features selected by Chi-square as it gave the highest accuracy as shown in Table 5. Table 7 shows the number of independent study documents which are classified into each group. As we observed that only one independent study document was classified as operational research in the year 2004 and 2005. We therefore investigated the title of independent study and found that there are no independent study researches in the field of operational research during the year 2004-2005. Thus, the classification model classified incorrectly, the reason is that the classification model is trained on a small training dataset and the selected features may inadequate for classification

Table 7 classification of the undergraduate independent Study in statistics during 2004-2009

Year	Group			Total
	Sample Survey	STAT Analysis	Operation Research and related field	
2004	9	9	1	19
2005	22	3	1	26
2006	25	10	4	39
2007	29	11	5	45
2008	34	17	5	56
2009	27	21	10	58

Conclusion and future work

This paper presents the classification of Thai undergraduate independent study in statistics using the data mining techniques. The RBF algorithm is selected for constructing the classification model since it gives the best results. And, the classification model is used to classify the titles of the independent study during the year 2004-2009 which have never been assigned the group label. However, there are many titles were misclassified into other groups. We investigated this problem and observed that the training dataset contains a small number research's titles. And, this could affect the accuracy of the model a lot. Future work will investigate the impact of skewed class distribution of the training dataset to the accuracy of the classification model. In addition, we will improve the accuracy of the classification model by considering other features such as advisor name, words in the abstract, etc. Moreover, we will conduct more experiments using clustering techniques, which is unsupervised learning method, in order to group the documents into more than three groups. The clustering results will show the best number of research groups for our department.

References

1. T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proc. of ECML-98, 10th European Conference on Machine Learning, Springer Verlag, Heidelberg, DE, 1998, 137-142.
2. D. D. Lewis and M. Ringuette, "Comparison of two learning algorithms for text categorization," Proc. Of SDAIR, 1994. 81-93.
3. N. Chirawichitchai, P. Sanguansat, and P. Meesad, "An Experimental Study on Feature Reduction Techniques and Classification Algorithms of Thai Documents," Journal of Science Ladkrabang, 2009. (in Thai)
4. N. Chirawichitchai, P. Sanguansat, and P. Meesad, "A comarative Study on Term Weight Techniques for Thai Document Categorization," Journal of Science Ladkrabang, 2010. (in Thai)
5. N. Chirawichitchai, "Automatic Thai Document Classification Model," The journal of Industrial Technology, Vol.9, No.1 January-April, 2013. (in press)
6. S. Meknavin and P. Charoenpornsawat, "Feature-based Thai Word Segmentation," Proc. of the Natural Language Processing Pacific Rim Symposium, 1997.

7. C. Jaruskulchai, An Automatic Indexing for Thai Text Retrieval. PhD Thesis, George Washington University, USA, 1998.
8. J. Han and M. Kamber, Data Mining Concepts and Techniques. Second Edition, Morgan Kaufmann, 2006, 348-349.
9. M. A. Hall, Correlation-based Feature Selection for Machine Learnin. PhD Thesis. Department of Computer Science, The university of Waikato, Newzealand, 1999.
10. Y. Yang and J. O. Pederson, “A Comparative Study on Feature Selection in Text Categorization,” Proc. of 14th International Conference on Machine Learning, 1997.
11. J. Platt, “Fast Training of Support Vector Machines using Sequential Minimal Optimization,” in B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, 1998.
12. S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. “Improvements to Platt’s SMO Algorithm for SVM Classifier Design,” Neural Computation, vol. 13(3), 2001, 637-649.
13. T. Hastie and R. Tibshirani, “Classification by Pairwise Coupling,” in Advances in Neural Information Processing Systems, 1998.
14. G. H. John and P. Langley, “Estimating Continuous Distributions in Bayesian Classifiers,” Proc. of the 11th Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995, 338-345.
15. J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA, 1993.
16. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA Data Mining Software: An Update,” SIGKDD Explorations, Vol. 11, Issue 1, 2009.