Journal of Theoretical and Applied Information Technology
10th August 2014. Vol. 66 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THE EFFECTIVENESS OF AUTOMATED THAI DOCUMENTS CATEGORIZATION BASED ON MACHINE LEARNING

[1]SATIEN JANPLA

[1] Computer Science, Science and Technology, Suan Sunandha Rajabhat University, Bangkok,THAILAND

E-mail: [1]satien@ssru.ac.th

ABSTRACT

The purpose of this research was to test the effectiveness of the Thai language document categorization model by using the features of reduction and machine learning techniques. The experimental results showed that the support vector machine algorithm used to classify Thai documents did so with a highest efficiency of 93.06%, followed by naïve-bayes and decision tree at 86.80% and 76.51% respectively. Considering the parameter that had the best efficiency, support vector machine algorithm with 7000 features had the best performance at 93.90%. The features could be reduced significantly without affecting the performance of categorization.

Keywords: *Model, Categorization, Machine Learning, Effectiveness, Automated*

1. INTRODUCTION

The quick expansion of computer systems, applications and the internet to the present day has resulted in the creation of many kinds of data in various electronic forms. These electronic information data types, such as electronic mail (E-mail) web page (Web page) news pages (News) File of documents (Document), which are diverse forms of information content, make it difficult to find and store the classification of documents. This is needed by professionals for grouping, but it is difficult to group or classify documents since the large volume of documents rises every day. The requirement for human resources in the classification of these documents rises every day as well. The development of a process of data classification with these large volumes is needed discriminate information in order to take advantage of the information and manage data effectively, supporting the retrieval of documents from the users correctly and appropriately.

Research needs to be focused on the classification of documents into groups based on the content. Attention needs to be focused on presenting techniques for grouping or categorizing and learning with computers in several ways, which these techniques can be applied to many types of data, texts, pictures and sounds by classifying documents. Those documents can be divided into 2 varieties Including grouping (Clustering) and classification (Classification or Categorization). Document clustering divides the group according to the contents of documents without defining a group or category of documents, which is divided according to the characteristics of the document. The documents which are similar are grouped together. The classification of the document is first based on segmenting the contents of the document, which is then assigned a group or category. This is then compared with the original documents in each category. The document will be arranged in a category that looks most similar to master category itself. The main idea of the document categorization is that it is done by means of Machine Learning [1],[2] which creates an automatic document classification (Automatic Text Classifier) with a computer by inductive learning methods from a set of documents that can classify beforehand. The characteristics of the relevant category, the advantages of the method of learning by computer, and the accuracy of the results are close to the classification of documents made by humans. This can save a lot of human labor, because it does not require an expert to identify the document classification or require someone to alter the document classification.[3]

The lack of research in the development of document classification in Thailand is because most of the current research into document categorization development focused on the development of English language document classification. Thai

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

language documents have a unique nature in the classification of documents, which is different than in English. In the Thai language there are issues found with consecutive strings of words and a lack of punctuation to show splits in sentences and sentence structure. These problems extracted from Thai documents are used in learning to create models for large volumes of document categorization. This requires system resources and a lot of time to process, such as with problems relating to the ambiguity of terms. As a result, there is a problem with the accuracy of document classification in Thai and the in effectively creating a highly effective index for indexing classifications

From the importance of these problems, this project was to develop a conceptual document classification system for Thai language documents to be used in learning to model large volume document classification that has many various features. This will be done by modeling document classification by having the computer learning the Thai language. The system is built to have the abilities to particularly classify automatically from any websites, news (criminal, economic, political, etc.), filtering emails so that it can extract only Thai language and management of Thai laws and regulations effectively.

2. THE PURPOSE OF THE RESEARCH

To evaluate the effectiveness of the automated Thai language document categorization by machine learning.

3. STEPS TO CATEGORIZE THAI DOCUMENTS

Based on artificial neural networks, in order to improve the efficiency of Thai language document categorization in accuracy of categories (Accuracy) and overall accuracy in retrieving information (F-Measure), 6 operational steps as depicted in figure 1 are undertaken as follow.

3.1 Data Collection

The actual data collected from Thai electronic documents and various online news services via the Internet was in 10 groups, including education, entertainment, social groups, political groups, technology groups, sports news, foreign, agricultural, economic, and Bangkok. In total, 12,000 documents were sent for the extraction processes.

3.2 Feature Extraction

After data collection was completed, the extraction feature was used because the computer could not identify the category of the document, which was directly because of natural language. Therefore, it had to convert the documents into a format that the computer could use, a process called feature extraction. The purpose of the important feature extraction is to extract attributes (Feature) of the documents involved, which pulled out features. First, we need to define what to use and what represents the characteristics of the document, using any features of value the documents have. The past research both domestically and abroad shows that features can be represented with a single word, syllables, phrases, groups of words, or sentences. They represent characteristics of the documents and the frequency of the words that appear in the document is the value of the feature.

Document feature extraction was done by bringing documents from various news-papers from websites and emails. Documents were converted to the same format, so there was an element of the documents that was the same. In this case it was that they were text documents (Plain text). It extracted the structural tag language design web away and instead used a single word, syllables, phrases, group of words, sentences, to represent the characteristics of the document. It then calculated the weight of the features. Feature extraction techniques may be used to help language analysis techniques (Language Analysis). It cuts out unnecessary words; such truncation is insignificant (Stop-word list removal). It makes root words (Word stemming) or cuts out unnecessary words, according to the archives, off of the main sentence. The main use of N-gram attenuation is applied to reduce the dimension of document size. [1],[2]

3.3 Word Segmentation

Processing the category of Thai language documents classification is done by learning with computers. It is done effectively in the initial processing by extracting the characteristics of the Thai language document. Truncation in Thai language is a necessary step (Word Segmentation). However, due to language problems, the Thai language has not symbols that indicate word boundaries, such as with English, which uses a space (Space) between the boundaries of words. This is one obstacle that needs to be crossed, to string language into words correctly. We developed a method of cutting the Thai word (Thai Word Segmentation), which searches on cutting of Thai

Journal of Theoretical and Applied Information Technology
10th August 2014. Vol. 66 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

words that are developed together in many ways. Each method has advantages and strengths of their own and gives different results in terms of accuracy, fast work and consumption of resources. Currently, the research and development of Thai language word segmentation has been continuous, which can be categorized according to the characteristics of the database used in word segmentation into 3 main groups: Rule-Based Approach, Dictionary-Based Approach, and Word Stemming Approach. [4],[5]

3.4 Indexing

The computer cannot distinguish the category of the document, because it cannot directly use natural language. To convert documents in a way computers can be used in the learning step, it creates an index (Indexing) instead of document content (Document Representation) for use in the learning process. Characteristics of document based on what you want to consider are used to determine the specific meaning of a word, words or what you would like to consider the meaning according to the rules of the language. This makes determinations based on the position of the words in the sentence. Classification by means of learning with a computer will determine the characteristics of representative documents use the popular meaning of a word, regardless of the position of words. It will ignore the representative meaning in documents, usually in the form of a weight vector. Important objectives in this section are indexed. These calculate the value of what is used as the attribute value of the document or it may be determined by the weight (Term weighting) which is a feature used in the form of a single word. [1],[2],[6],[7],[8]

3.5 Model

Process modeling needs to rely on the algorithms in the classification for learning effective solutions (Supervised Learning). The algorithm can be divided into document categorization. It is divided into 2 steps: learning to build a template and separate categories of documents of interest, by examining the underlying documents similar to the group. The process is used to find images of the series that are similar or the most common. This is used to predict the data set that in any type of data set was split. The data set is divided into a set of learning from the existing series model (Training data) that is already there. In this paper, models resulting from learning. [9],[10]

3.6 Performance Model

Performance evaluation assesses the ability of the model to the classification of Thai language documents by learning with computers that emphasized the ability to make decisions or predictions on the correct category. The method is tested for com-paring the efficiency of the development model to the efficiency of document classification. This can be determined by the accuracy, by measuring the effectiveness of information based on information retrieval, which is the measurement accuracy , Precision, Recall and the efficiency of the F-Measure by the developing stage must be effective in classification of documents, no less than 80%. [9],[10]

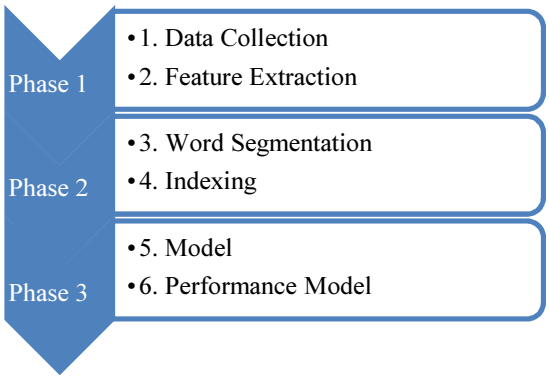

Figure 1: Steps In Automated Thai Language Document Categorization

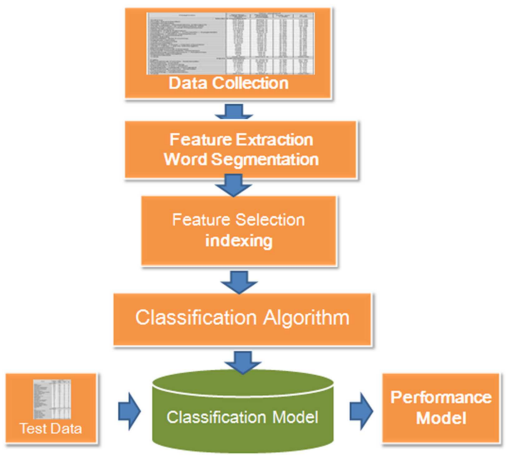

Figure 2: Model Automated Thai Language Document Categorization

4. EXPERIMENTAL RESULT

The experiment carried out demonstrated the evaluated model can be applied to the task of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

categorizing web pages, classification of site groups, and classification of newsgroups, classification file documents to search and sorting email. It was tested with electronic news from newspapers online. The experiment tested ways to reduce the feature value method of gaining information (Information gain) together with machine learning, which consisted of a support vector machine algorithm (SVM), decision trees (DT) and Naïve Bayes (NB). The conclusion is as follows.

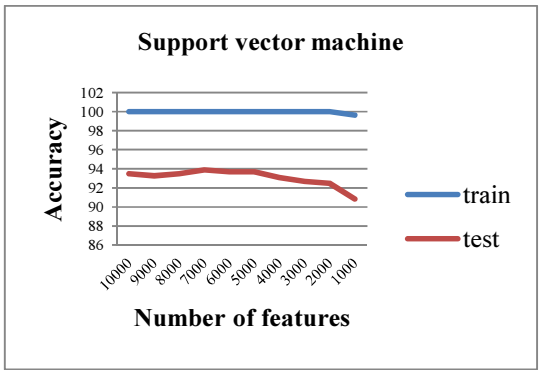

Figure 3: Graph Compares The Algorithm Support Vector Machine.

From the experiment the weight to the index of a TFIDF-weighting reduced feature with information gain value (Information Gain) and learning with a support vector machine algorithm (SVM) was measured in performance from the accuracy (Accuracy). It can be concluded that the model was created. The document categorization was effective. The average in the study group (Training set) was 99.96% and the average in the test group (Test set) was 93.06%. When considering the parameters, the results in efficiency of document categorization with support vector machine were best. The result found for the number of 7000 features, that the performance level was 93.90% as best Figure 3.

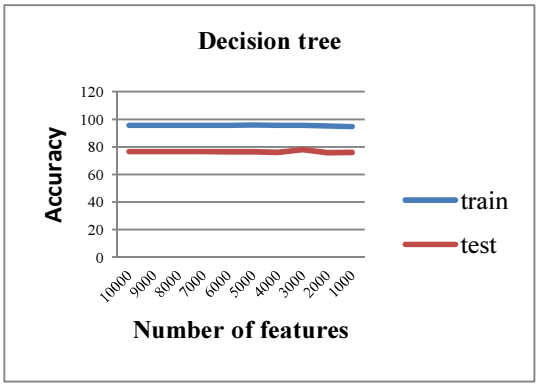

Figure 4: Graph Compares Decision Tree Algorithm.

From the experiment, the weight to the index of a TFIDF-weighting reduced feature with information gain value (Information Gain) with learning decision tree algorithm (Decision tree) was measured by the performance from the accuracy (Accuracy). It can be concluded that the model was created. The document categorization was effective. The average in the study group (Training set) was 95.49% and average in the test group (Test set) was 76.51%. When considering the parameters, the best result in efficiency of the document categorization decision tree, was found with the number of 3000 features, and the performance level was 78.04% as in Figure 4.

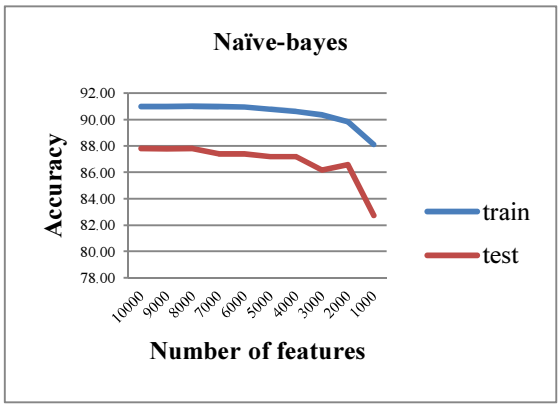

Figure 5: Graph Performance Comparison Of Algorithms With Naïve Bayes

From the experiment, the weight to the index of a TFIDF-weighting reduced fea-ture with information gain value (Information Gain) with learning algorithms, Naïve Bayes was measured by performance from the accuracy (Accuracy). It can be con-cluded that the model was created. The document categorization was effective. The average in the study group (Training set) was 90.46% and the average in the test (Test set) was 86.80%. When considering the parameters, the result in best efficiency of document categorization, was found for the number of 8000 features, the performance level was 87.80% as in Figure 5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

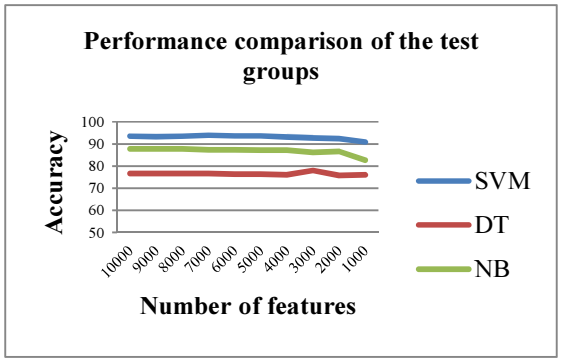

Figure 6: Graph Comparison Performance Of Each Algorithm..

From the experiment, the weight to the index of a TFIDF -weighting reduced feature with information gain value (Information Gain) with learning support vector machine (SVM), decision trees (DT), algorithm and Naïve Bayes (NB) to measure efficiency of the accuracy (Accuracy) concluded the classifier support vector machine algorithm performance in document categorization was best followed by a 93.06% algorithm Naïve Bayes. The performance in the classification of documents was 86.80% and finally, the decision tree algorithm was effective in the classification of documents at 76.51%, as shown in the Figure 6.

5. CONCLUSION AND DISCUSTION

Summing the results of the study showed the main factors that affect the efficiency in the evaluation of an automated Thai document classification Based on feature extraction methods which were appropriate and how to keep the weight of terms, which was used in this research method. TFIDF -weighting is a standard and has been very popular. The other factors that will contribute to the efficiency of the higher classification were caused by behavioral factors, and the distribution of the sample in Thailand. Factors and the frequency of words that appear in documents (Term frequency) affect the performance of document classification because of the idea that the more occurrences of that word in the document, the better. Terms are more likely to be associated with the clustering of documents and are even better. The results showed that the frequency of words appearing in each case is an important factor for a high efficiency in document classification. What is being done differently in this paper from the others is that various algorithms are being compared for their efficiencies along with adjusting the relevant parameters particularly the dimensional reduction of the data based on Thai language corpora in order that the optimal solution of the efficiency in automating the classification of Thai documents are reached.

REFRENCES:

[1] Nichaporn Sura. 2006, "Classification automatic Thai document using an algorithm FPTC", *Thesis Master of Science In Computer Science*,King Mongkut's University of Technology north Bangkok.

[2] Wallop In-Chum. 2005, "Automatic Thai document categorization system using SVM with processing languages", *Thesis Master of Engineering in computer*, Kasetsart University.

[3] Fabrizio Sebastiani, "Machine Learning in Automated Text Categorization", *ACM Computing Surveys*, Vol. 34, No. 1, March 2002, pp. 1–47.

[4] Yuen Poovarawan and Wiwat Im-Arrom, 1986, "Thai syllable separator by dictionary", *Conference on electrical engineering, 9th*.

[5] Wirath Sornleanlomwanich, 1993, "To Thai word segmentation in machine translation system Machine translation", *National Electronics and computer technology center*, Bangkok.

[6] Aas, and Eikvil, 1999, "Text Categorization: a Survey", *Report No. 94. Norwegian Computing Center*.

[7] Jaruskulchai, 1998, "An Automatic Indexing for Thai Text Retrieval", *PhD Thesis, George Washington University*, USA.

[8] Yang and Perderson, 1997, "A Comparative Study on Feature Selection in Text Categorization", *Proceedings of ICML-97, 14th International Conference on Machine Learning*.

[9] Han and Kamber, 2006, "Data Mining Concepts and Techniques", *San Francis-co. Morgan Kaufmann Publishers*.

[10] Ian H. Witten, Eibe Frank, 2005, "Data mining : practical machine learning tools and techniques", Amsterdam.